5-2016

Separation of Points and Interval Estimation in Mixed Dose-Response Curves with Selective Component Labeling

Darl D. Flake II
Utah State University

SEPARATION OF POINTS AND INTERVAL ESTIMATION IN MIXED DOSE-RESPONSE CURVES WITH SELECTIVE COMPONENT LABELING

by

Darl D. Flake II

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

DOCTOR OF PHILOSOPHY

in

Mathematical Sciences
(Statistics)

Approved:

John R. Stevens Major Professor	Daniel C. Coster Committee Member
Adele Cutler Committee Member	Edward W. Evans Committee Member
Guifang Fu Committee Member	Mark R. McLellan Vice President for Research and Dean of the School of Graduate Studies

UTAH STATE UNIVERSITY
Logan, Utah

2016

ABSTRACT

Separation of Points and Interval Estimation in Mixed Dose-Response Curves with Selective Component Labeling

by

Darl D. Flake II, Doctor of Philosophy

Utah State University, 2016

Major Professor: John R. Stevens
Department: Mathematics and Statistics

This dissertation develops, applies, and investigates new methods to improve the analysis of logistic regression mixture models. An interesting dose-response experiment was previously carried out on a mixed population, in which the class membership of only a subset of subjects (survivors) were subsequently labeled. In early analyses of the dataset, challenges with separation of points and asymmetric confidence intervals were encountered. This dissertation extends the previous analyses by characterizing the model in terms of a mixture of penalized (Firth) logistic regressions and developing methods for constructing profile likelihood-based confidence and inverse intervals, and confidence bands in the context of such a model. The proposed methods are applied to the motivating dataset and another related dataset, resulting in improved inference on model parameters. Additionally, a simulation experiment is carried out to further illustrate the benefits of the proposed methods and to begin to explore better designs for future studies. The penalized model is shown to be less biased than the traditional model and profile likelihood-based intervals are shown to have better coverage probability than Wald-type intervals. Some limitations, extensions, and alternatives to the proposed methods are discussed.

(82 pages)

PUBLIC ABSTRACT

Separation of Points and Interval Estimation in Mixed Dose-Response Curves with Selective Component Labeling

by

Darl D. Flake II, Doctor of Philosophy

Utah State University, 2016

Major Professor: John R. Stevens
Department: Mathematics and Statistics

Dose-response experiments are those that involve giving subjects different amounts of a treatment and observing the outcome. For example, plants may be given fertilizer and their growth could be measured or cancer patients could be given different doses of chemotherapy and their response could be monitored. These experiments are used to understand the relationship between the amount of, and response to, the treatment. Logistic regression models are often used to summarize data from these types of experiments. The dose-response experiment that motivated this dissertation involved treating a grain-pest with a pesticide. Some of the beetles had genes that made them more sensitive to the pesticide. However, the genes were only looked for in the beetles that survived the treatment. Additionally, traditional statistical models yielded unreliable results when they were applied to this data. Both specific summary values (parameter estimates) and likely ranges of values (confidence intervals) were not reasonable. This dissertation developed new statistical methods to improve the statistical modeling of dose-response experiments like this one. Two methods that are used in simpler situations, were applied to this dataset to overcome these problems: a Firth penalty and profile likelihood-based confidence intervals. The Firth penalty improved the parameter estimates and the profile likelihood-based confidence intervals were an improvement over the traditional confidence

intervals. Simulations were used to show that proposed methods worked well in a variety of situations. The statistical methods developed here are applicable to other situations not limited to dose-response experiments.

For Tiffany, through it *all*.

ACKNOWLEDGMENTS

Returning to school and, specifically, undertaking this research has been inextricably connected with the other major aspects of my life, which predominantly have been family, work, and church service. I have received so much support during this endeavor, for which, I can never fully express my gratitude. Regardless, I will make a feeble attempt.

First, and foremost, I'd like to thank my sweet wife, Tiffany, whose chosen field of research is not only more important but also fraught with more setbacks than any other. My children Kayden, Riley, and Sam have given me hope and brought joy to my life. John Stevens has allowed me to suffer through my independence but has been available on demand. I'm grateful that he was willing to guide me along this non-traditional path. Sasha Gutin, the smartest person I know, has taught me that simpler is better and that understanding is crucial. Julia Reid has been incredibly patient with me even when I had a hard time believing her when she pointed out that the world does not revolve around me. Susanne Wagner for being a scientist in the truest sense of the word and a caring individual. She probably doesn't think the title of this dissertation is nearly complicated enough. All the members and leaders of my local church congregations for not judging me for falling short in my responsibilities even when they probably had no idea why. Above all, I am grateful to God who has "enlighten[ed] my understanding" (Alma 32:28). "I know in whom I have trusted." (2 Nephi 4:19)

Darl D. Flake II

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Schlipalius et al. [1] first described the experiment that motivated this dissertation. The researchers carried out a genetic linkage study to identify chromosomal regions that are responsible for phosphine resistance in *Rhyzopertha dominica*, a stored grain pest. Two regions, rp5.11 and rp6.79, were identified as having a significant association with resistance to the insecticide. A sample of 10,798 beetles from the F_5 generation, containing all six combinations of genotypes, were used in a dose-response experiment in which the beetles were treated with one of 11 doses of phosphine. To conserve resources, only the 378 surviving beetles were genotyped for the two genes. The primary goal of the researchers was to reconstruct dose-response curves for each of the genotypes while accounting for the missing genotypes of the beetles that were killed by the insecticide in order to estimate the lowest dose necessary to kill a large proportion of the treated beetles (e.g. LD_{99}, the dose that is lethal to 99% of subjects). The design and resulting data for this experiment are shown in Table 1.1.

1.1 Logistic Regression Modeling of Dose-Response

Logistic regression has been used to model dose-response data for decades (e.g. Berkson [3]). While maximum likelihood estimation is generally used to fit logistic regression models, direct maximization of the likelihood is not straightforward for the beetle data in Table 1.1, due to the missing genotypes of the dead beetles. However, the expectation-maximization (EM) algorithm [4] is a general purpose algorithm for maximum likelihood estimation where incomplete data are present or where the likelihood can be expressed in terms of a latent variable. The EM algorithm maximizes the likelihood by iteratively maximizing a surrogate function, the expected conditional log-likelihood:

$$Q(\theta \mid \theta') = E(\mathscr{L}(\theta \mid x) \mid y, \theta'), \tag{1.1}$$

where θ are the parameters of the log-likelihood $\mathscr{L}$, θ' is the proposed or assumed value of the

Table 1.1: Data for a dose-response experiment from Schlipalius et al. [1] in which *Rhyzopertha dominica* were treated with phosphine.

Outcome (genotype)	Phosphine Dosage ($\mu g/L$)											Total
	0	3	4	5	10	50	100	200	300	400	1000	
Dead (NA)	0	16	68	78	77	270	383	740	490	492	7806	10420
Alive (-/B)	31	18	10	1	0	0	0	0	0	0	0	60
Alive (-/H)	27	26	4	4	1	0	0	0	0	0	0	62
Alive (-/A)	10	10	3	7	9	0	0	0	0	0	0	39
Alive (+/B)	6	6	4	2	8	5	0	0	0	0	0	31
Alive (+/H)	20	20	7	6	5	20	10	0	0	0	0	88
Alive (+/A)	4	4	4	2	0	5	7	10	10	8	44	98
Total	98	100	100	100	100	300	400	750	500	500	7850	10798

parameters (usually the estimate of θ at the previous iteration denoted as $\hat{\theta}^{(k)}$), and y and x are the observed and unobserved data, respectively. Ibrahim [5] worked out the details of the EM algorithm in the context of generalized linear models, facilitating the fitting of logistic regression models with missing values.

Stevens and Schlipalius [6] first analyzed the beetle data using an implementation of the EM algorithm to obtain parameter estimates from the maximized likelihood. They estimated the asymptotic variance-covariance matrix of the parameter estimates by direct calculation of the information matrix [7]. The estimated parameters were used to calculate median lethal doses (LD_{50}) and corresponding 95% inverse and Fieller intervals. The authors noted that for genotypes -/A and +/B the fitted dose-response curves appeared to "jump" steeply, either between, or at observed doses. Additionally, the variance estimates for the dose-response curve parameters for these genotypes were extremely large. As a result, the inverse intervals were large and the Fieller intervals did not exist.

1.2 Separation of Points, Firth Logistic Regression, and Bias Reduction

In a subsequent analysis of the beetle data by Rounds [8], he noted that complete or quasi-complete separation of the data was the likely cause of the inflated standard error estimates of Stevens and Schlipalius [6]. Complete separation of points occurs when the predicted probability of all events is greater than the predicted probability of every non-event. Quasi-complete separation of points occurs when the predicted probability of every event is greater than or equal to the predicted

probability of every non-event. When such separation is present, maximum likelihood estimates of parameters do not exist or are not unique [9]. When fitting models to data with separation, predicted probabilities numerically approach 0 or 1 and the likelihood converges while the parameter estimates do not.

The model fit in Stevens and Schlipalius [6] demonstrates examples of both types of separation. The predicted probability of death for beetles with genotype -/A was (practically) zero for all doses up to and including 10 $\mu g/L$ of phosphine which also correspond to doses were surviving beetles from genotype -/A were observed (see Table 1.1). No surviving -/A beetles were observed at the next higher dose, 50 $\mu g/L$, of phosphine. Therefore, all surviving beetles are assumed to have received lower doses than all dead beetles for the genotype -/A (i.e. complete separation of points). In the case of genotype +/B, the predicted probability of death was zero for all observed doses up to and including 10 $\mu g/L$ of phosphine. Some surviving +/B beetles were observed at 50 $\mu g/L$ and the predicted probability of death was approximately 0.69. At the next observed dose, 100 $\mu g/L$, there were no observed surviving beetles with genotype +/B and the predicted probability of death was one. Therefore, all surviving +/B beetles received doses lower than or equal to those received by all dead +/B beetles (i.e. quasi-complete separation of points).

Separation of points can be viewed as an extreme case of another problem: bias in logistic regression estimates. With finite samples, the maximum likelihood estimate of the slope term from logistic regression is over-estimated, resulting in predicted probabilities closer to 0 and 1 than the truth. The more extreme the predicted risk, the larger the bias becomes. In the case of complete separation of points, all observations have extreme risk and the estimates are biased towards infinity.

Firth [10] introduced a penalty on the likelihood with the primary goal of reducing bias in maximum likelihood estimates from generalized linear models, including logistic regression. Heinze and Schemper [11] showed that this same penalty is a good solution to the problem of separation of points. In addition to reducing bias, another convenient side effect of the penalty term in logistic regression is that it also reduces the estimated standard error of the parameters. The penalty term forces the estimated probabilities closer to 0.5 compared to the traditional model. Due to the relationship between the mean and variance of a binomial random variable, the standard error estimates

are also smaller. While the Firth penalty has predominantly been used as a remedy for separation of points, it could argued that this penalty should be used in all logistic regression models.

Dempster et al. [4] stated and Green [12] later proved that including a penalty term G in Q during the "M" step of the EM algorithm yields the maximum penalized likelihood estimate:

$$Q_{pen}(\theta \mid \theta') = Q(\theta \mid \theta') + G(\theta) \tag{1.2}$$

Rounds [8] proposed a penalized EM algorithm to overcome the challenges encountered by Stevens and Schlipalius [6] in their analysis of the beetle data. Instead of maximizing the standard binomial likelihood from logistic regression, he maximized the likelihood with a Firth penalty. While unnecessary, in the "E" step of the penalized EM algorithm, he planned to take the expectation of the conditional log likelihood with the penalty:

$$Q_{pen}(\theta \mid \theta') = E(\mathscr{L}(\theta \mid x) + G(\theta) \mid y, \theta'), \tag{1.3}$$

He found this to be analytically intractable because the penalty was a function of the missing data. Therefore, a modified estimate of the expected value was used in place of the true expected value without sufficient justification. The modification was to use the parameter estimates from the previous iteration for the penalty term, inadvertently making it equivalent to the proposed method of Dempster et al. [4].

1.3 Mixtures of Logistic Regression Models

Stevens and Schlipalius [6] and Rounds [8] viewed the beetle data as an application of logistic regression with missing categorical covariates. The logistic regression model they chose included interaction terms between genotype and both the intercept and slope terms of the logistic regression model. In the event of no missing data, this is no different than fitting individual dose-response models to each genotype. Wang and Puterman [13] described the EM algorithm for the special case of logistic regression models where the missing data were the latent class membership of each (and every) observation. Redner and Walker [14] identified different types of mixture models that

they distinguished by the extent of partial observation of the class membership in the data. Thus, a different way to characterize the beetle data is in terms of a mixture model where component membership is observed for a subset of the subjects. Following the framework of Redner and Walker [14], the beetle data can be modeled as a Type 4 mixture of logistic regressions, which is presented later in Equation 2.4.

1.4 Confidence, Prediction, and Inverse Intervals

Confidence intervals, prediction intervals, and inverse intervals are traditionally derived from the asymptotic variance-covariance matrix of the parameters [15]. This is done by calculating either the observed, or expected (Fisher's), information matrix and negating its inverse. Variance estimates for the raw parameters are taken directly from the diagonal elements of the variance-covariance matrix. Variance estimates for predicted values and their corresponding doses can be calculated by applying the delta-method to the variance-covariance matrix. The sampling distribution for the parameter estimates is assumed to be normal, and confidence intervals are chosen to match the corresponding quantiles. The resulting Wald-type intervals are symmetric about the estimated quantity. Additionally, Fieller intervals, asymmetric inverse intervals based on the distribution of the ratio of the intercept relative to the slope, can also be calculated using the asymptotic variance-covariance matrix [16]. While Fieller intervals are preferred over Wald-type inverse intervals, they do not always exist.

In their analyses of the beetle data, Stevens and Schlipalius [6] and Rounds [8] calculated traditional Wald-type confidence intervals for the parameter estimates. As mentioned above, they calculated the asymptotic variance-covariance matrix via the methods of Oakes [7]. Stevens and Schlipalius [6] also used the asymptotic variance-covariance matrix from the traditional EM algorithm together with the delta method to construct confidence and inverse intervals on the predicted mortality from the dose-response curves. Fieller intervals were constructed but did not exist for the two genotypes that displayed steep dose-response curves. While Rounds [8] did not calculate inverse or prediction intervals, he did carry out a small simulation study to evaluate the coverage probability of the confidence intervals on the parameters. He observed that reliance on the asymp-

totic normality of the sampling distribution of the parameters was not entirely justified. Specifically, the coverage probability of the Wald-type confidence intervals on the model parameters was low.

Hudson [17] proposed an alternative to Wald's method for constructing confidence intervals of single parameter distributions that he called likelihood intervals. His method consisted of "drawing a horizontal line across the graph of the likelihood function." The value of the horizontal line is generally chosen to correspond to the critical value of the likelihood ratio test. The method for getting likelihood intervals can be generalized to probability distributions involving more than one parameter by likelihood profiling. While called "maximum relative likelihood" by Kalbfleish and Sprott [18], and "the maximized likelihood" by Patefield [19], the "profile likelihood" is the likelihood function of a parameter of interest, calculated by optimizing over all other (nuisance) parameters.

Like Wald-type confidence intervals, profile likelihood-based confidence intervals are also approximate. However, because profile likelihood-based confidence intervals depend on the asymptotic distribution of the likelihood ratio, they are often preferred over Wald-type confidence intervals which rely on the asymptotic normality of the individual parameter estimates [20]. This may be particularly true when the sample size is small or when an estimate is near a boundary of the parameter space [21] as is the case with the beetle data. Specifically, Heinze and Schemper [11] showed that profile likelihood-based confidence intervals are often preferred for penalized (Firth) logistic regression. Even with the penalized model, such intervals are likely to suffer from the small effective sample size of some of the less abundant mixture components. Profile likelihood-based confidence intervals are also invariant to reparameterization [21, 22].

Profile likelihood-based confidence bands have been calculated, but only in the absence of missing data. Based on earlier work (see Bjørnstad [23] for a review), Kreutz et al. [24] explored the prediction profile likelihood and obtained confidence bands on predictions from complicated models. To obtain confidence bands based on likelihood profiling, one must maximize the likelihood, subject to the constraint that the parameters yield a predicted value of interest. If the constrained maximum likelihood is significantly lower than the unconstrained maximum likelihood, then the predicted value of interest is outside of the profile likelihood-based confidence interval. A grid

search on predicted values can be carried out to find the endpoints of the interval.

Similarly, profile likelihood-based inverse intervals have been calculated in the absence of missing data. For example, Williams [25] proposed calculating confidence intervals on LD_{50} using likelihood profiling and Alho and Valtonen [26] extended these results to include inverse intervals for dose response models with additional covariates and arbitrary predicted risk. For an inverse interval on LD_{50}, the likelihood is maximized subject to the constraints that the dose is fixed to yield a predicted probability of 50%. If the fixed dose value yields a constrained maximum likelihood that is significantly different from the unconstrained maximum likelihood, then it is outside the profile likelihood-based inverse interval on LD_{50}. Profile likelihood-based inverse intervals may have lower coverage probability compared to Fieller intervals, but they always exist [26].

Obtaining profile likelihood confidence, inverse, and prediction intervals requires constrained optimization. Because logistic regression is a generalized linear model, only linear constraints of the parameters are needed for likelihood profiling of predicted probabilities and their corresponding inverse intervals. Kim and Taylor [27] demonstrated that, using the restricted EM algorithm, one can maximize likelihood functions with missing values and linear constraints on the parameters. As an example they found profile likelihood-based confidence intervals on the parameters of a model with missing data within the context of the EM algorithm.

1.5 Outline

Various challenges were encountered in previous analyses of the beetle data; although, significant progress has been made. This dissertation will extend these methods to provide better inference on, and interpretation of, the quantities of interest. In chapter 2, the methodological details for the implementation of a mixture of penalized (Firth) logistic regressions and the construction of the corresponding Wald-type and profile likelihood-based confidence, prediction, and inverse intervals will be described. In chapter 3, the developed methods will be applied to the beetle dataset and another similar real dataset. Chapter 4 focuses on a simulation study of different scenarios to improve the experimental approach of the previous dosing studies in which the bias of the Firth logistic regression mixture model and the coverage probabilities of the confidence and inverse intervals will

be assessed. Chapter 5 concludes the dissertation with a discussion of the benefits, limitations, and extensions of, and the alternatives to, the described methods.

CHAPTER 2

MIXED FIRTH LOGISTIC REGRESSION

This chapter develops the proposed model in detail and describes the algorithm that can be used to fit it. Similarly, the various confidence intervals and methods to calculate them are explained.

2.1 Model Details

Let y_{ij} be a realization from a binary random variable whose generating distribution is a mixture of m logistic regression models. We assume that the relationship between the probability of an event for observation j of class i with covariate vector x_{ij} follows the formula,

$$\pi_{ij} = \frac{1}{1 + exp{-x_{ij}\beta_i}}, \tag{2.1}$$

where β_i is a vector of logistic regression parameters, for class i.

The corresponding probability density function for a single class is

$$p_i(y_{ij} \mid \beta_i) = (\pi_{ij})^{y_{ij}} (1 - \pi_{ij})^{1-y_{ij}}, \tag{2.2}$$

where y_{ij} is the binary event outcome for subject j from class i.

Following Redner and Walker [14], and letting α be the class mixture probabilities, we define

$$p(y_{ij} \mid \alpha, \beta) = \sum_{i=1}^{m} \alpha_i p_i(x_{ij} \mid \beta_i). \tag{2.3}$$

Then, the log-likelihood of this Type 4 mixture model [14] is:

$$\mathscr{L}(\alpha, \beta) = \sum_{j=1}^{N_0} \log p(y_{0j} \mid \alpha, \beta) + \sum_{i=1}^{m} \sum_{j=1}^{N_i} \log \alpha_i p_i(y_{ij} \mid \beta_i) + \log \frac{N!}{N_0! \cdots N_m!}, \tag{2.4}$$

where N_i is the number of subjects from each class i and class $i = 0$ refers to unlabeled subjects.

2.2 Estimation via the EM Algorithm

This section describes the details for implementing an EM algorithm for a mixture of penalized firth logistic regression models.

2.2.1 Initialization

In practice, multiple sets of starting values for the parameters being fit by the EM should be used to ensure that a global maximum is achieved. However, it is convenient to initialize β at zero and α uniformly across classes, and this approach appears to function well in practice (e.g. Chapter 3).

2.2.2 Expectation

The expected conditional log-likelihood function for an unpenalized mixture is

$$Q(\alpha,\beta \mid \alpha',\beta') = \sum_{i=1}^{m}\left[N_i + \sum_{j=1}^{N_0}\frac{\alpha_i' p_i(y_{0j} \mid \beta_i')}{p(y_{0j} \mid \alpha',\beta')}\right]\log\alpha_i + \sum_{i=1}^{m}\left[\sum_{j=1}^{N_i}\log p_i(y_{ij} \mid \beta_i) + \sum_{j=1}^{N_0}\log p_i(y_{0j} \mid \beta_i)\frac{\alpha_i' p_i(y_{0j} \mid \beta_i')}{p(y_{0j} \mid \alpha',\beta')}\right]. \quad (2.5)$$

Following Dempster et al. [4] and Green [12], for penalized maximization, the expected conditional log-likelihood can be replaced with

$$Q_{pen} = Q + \sum_{i=1}^{m} J_i \quad (2.6)$$

where the Firth penalty, J_i, is

$$J_i = \frac{1}{2}\left|X^T \operatorname{diag}\{W\} X\right|, \quad (2.7)$$

$$X = \begin{bmatrix} X_0 \\ X_i \end{bmatrix}, \quad (2.8)$$

where X_i is a matrix with rows comprised of x_{ij}, for $j = 1,\ldots,N_i$, and

$$W = \begin{bmatrix} W_0 \\ W_i \end{bmatrix}, \tag{2.9}$$

where W_0 and W_i are vectors comprised, in order, of

$$W_{0j} = \frac{\alpha_i' p_i(y_j \mid \phi_i')}{p(y_j \mid \Phi')}\pi_j(1-\pi_j), \tag{2.10}$$

for $j = 1,\ldots,N_0$, and $W_{ij} = 1$, for $j = 1,\ldots,N_0$.

2.2.3 Maximization

At each "M" iteration of the algorithm, the mixing probabilities, α_i, are maximized directly in the following way:

$$\alpha_i^+ = \frac{1}{N_i}\sum_{k=1}^{N_i}\frac{\alpha_i' p_i(y_{0j} \mid \beta_i')}{p(y_{0j} \mid \alpha', \beta')}. \tag{2.11}$$

This is not different than the "M" step for the mixing probability of an unpenalized logistic regression model because α does not appear in J.

The logistic regression parameters, β_i, are found according to

$$\beta_i^+ \in \arg\max_{\beta_i} \sum_{j=1}^{N_i} \log p_i(y_{ij} \mid \beta_i) + \sum_{j=1}^{N_0} \log p_i(y_{0j} \mid \beta_i)\frac{\alpha_i' p_i(y_{0j} \mid \beta_i')}{p(y_{0j} \mid \alpha', \beta')} + J_i. \tag{2.12}$$

This is just a Firth logistic regression with weighted observations [8]. This maximization can be carried out by the Newton-Raphson method described in Heinze and Schemper [11] and implemented in any software that performs Firth logistic regression. Then the updated vector of parameter proposals for the next iteration is $\theta' = (\alpha^+, \beta^+)$.

2.2.4 Convergence

The algorithm is repeated until the differences in subsequent θ are less than some small value (e.g. 1×10^{-5} in the application in Chapter 3).

2.3 Asymptotic Variance-Covariance Matrix

Following Oakes [7], the observed information matrix of the model parameters is a function of the sum of two matrices of second partial derivatives of Q:

$$-I(\theta) = \frac{\partial^2 \mathscr{L}(\theta, y)}{\partial \theta^2} = \left\{ \frac{\partial^2 Q(\theta \mid \theta')}{\partial \theta^2} + \frac{\partial^2 Q(\theta \mid \theta')}{\partial \theta \theta'} \right\}_{\theta'=\theta}. \tag{2.13}$$

The asymptotic variance-covariance matrix Σ is then the inverse of the observed information. That is, $\text{Var}(\theta) = [-I(\theta)]^{-1}$. While the matrix involves multiple derivatives it can be calculated computationally via a computer algebra system like Maple. As mentioned in Firth [10] the first order approximation of the asymptotic variance-covariance matrix for the penalized logistic regression model is the same as that of the traditional logistic regression model in that it excludes the penalty term. Therefore the penalty is excluded here as well.

2.4 Intervals Based on Asymptotic Variance-Covariance Matrix

In this section details for calculating confidence and inverse intervals and confidence bands based on the asymptotic variance-covariance matrix will be described.

2.4.1 Confidence Intervals on Parameter Estimates

Where $\Sigma = \text{Var}(\hat{\theta})$, a $1-\alpha$ Wald-type confidence interval on an individual parameter θ_l is calculated straightforwardly as $\hat{\theta}_l \pm z_{1-\frac{\alpha}{2}} \Sigma_{ll}$, where σ_{ll} is the diagonal element of Σ corresponding to θ_l.

2.4.2 Confidence Bands on Predicted Probability

Let β_0 and β_1 be the, respective, intercept and slope from a single-class logistic regression model. The most appropriate scale for normality of the predicted value is that of the linear predictor or dose:

$$LD_\pi = \frac{\log\left(\frac{\pi}{1-\pi}\right) - \hat{\beta}_0}{\hat{\beta}_1}, \tag{2.14}$$

where LD_π is the dose needed to achieve predicted probability π.

Therefore a $1-\alpha\%$ confidence band for π can be calculated in the following way:

$$\text{logit}^{-1}\left[LD_\pi \pm z_{1-\frac{\alpha}{2}}\begin{pmatrix}1 & LD_\pi\end{pmatrix}\Sigma_{\beta_0\beta_1}\begin{pmatrix}1\\ LD_\pi\end{pmatrix}\right], \tag{2.15}$$

where $\Sigma_{\beta_0\beta_1}$ is the portion of $\Sigma = \text{Var}(\hat{\theta})$ corresponding to β_0 and β_1, and $\text{logit}^{-1}(v) = (1+e^{-}v)^{-}1$.

2.4.3 Inverse Intervals

Following Stevens and Schlipalius [6] and by the delta method, the first order approximation of the Wald-type inverse interval for LD_π is:

$$LD_\pi \pm z_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{\beta_1^2}\begin{pmatrix}1 & LD_\pi\end{pmatrix}\Sigma_{\beta_0\beta_1}\begin{pmatrix}1\\ LD_\pi\end{pmatrix}} \tag{2.16}$$

2.4.4 Fieller Intervals

A $1-\alpha\%$ Fieller interval [16] for LD_π can be calculated in the following way:

$$-\frac{\left(\hat{\beta_0}^*\hat{\beta}_1 - \chi^2_{1-\alpha}\sigma_{\hat{\beta_0}^*\hat{\beta}_1}\right) \pm \sqrt{(\hat{\beta_0}^*\hat{\beta}_1 - \chi^2_{1-\alpha}\sigma_{\hat{\beta_0}^*\hat{\beta}_1})^2 - (\hat{\beta_0}^{*2} - \chi^2_{1-\alpha}\sigma_{\hat{\beta_0}^*\hat{\beta_0}^*})(\hat{\beta_1}^2 - \chi^2_{1-\alpha}\sigma_{\hat{\beta}_1\hat{\beta}_1})}}{\hat{\beta_1}^2 - \chi^2_{1-\alpha}\sigma_{\hat{\beta}_1\hat{\beta}_1}}, \tag{2.17}$$

where $\hat{\beta}_0^* = \log(\frac{\pi}{1-\pi}) - \hat{\beta}_0$ and $\sigma_{ab} = \text{Cov(a, b)}$.

2.5 Profile Likelihood-Based Intervals

In the context of the EM algorithm there is no closed form for the profile likelihood. Therefore, in order to evaluate a single point on a profile likelihood one must be able to maximize the likelihood subject to a constraint on the parameter(s) of interest. In the case of logistic regression and the types of confidence intervals to be calculated in this dissertation, only linear constraints are needed. In practice, a constraint on a likelihood is no different than a penalty within the EM algorithm. Therefore, like the Firth penalty, linear constraints can be applied to Q during the "M" step and the desired result will be obtained.

2.5.1 Linear Constraints within the EM Algorithm

Kim and Taylor [27] described a method for applying linear constraints to parameters within the EM algorithm. Such constraints can be used to carry out likelihood ratio based hypothesis tests based on linear combinations of parameter estimates or to profile the likelihood with respect to parameters of interest.

Following the methods of Kim and Taylor [27], let

$$\mathbf{A}\theta = \mathbf{a} \tag{2.18}$$

represent a linear constraint on the parameter vector θ. The constrained maximum likelihood estimate is found by calculating the restricted estimate from the unrestricted estimate at each iteration of the EM algorithm:

$$\theta_R(l+1) = \theta_U(l+1)[\theta_R(l)] + \mathbf{I}_U^{-1}\mathbf{A}^T(\mathbf{A}\mathbf{I}_U^{-1}\mathbf{A}^T)^{-1}(\mathbf{a} - \mathbf{A}\theta_U(l+1)[\theta_R(l)], \tag{2.19}$$

where $\theta_R(l)$ and $\theta_U(l)$ are, respectively, the restricted and unrestricted estimates and $\mathbf{I}_U$ the corresponding information matrix at iteration l. In the context of the current study, the unrestricted estimates are those found by penalized (Firth) logistic regression for the given "M" step.

In order to calculate the different types of intervals of interest, one only needs to change the constraint in Equation 2.18.

2.5.2 Confidence Intervals on Parameters

In the "M" step of the EM algorithm we apply Kim and Taylor [27]. The constraint required is that the single parameter of interest be fixed. Consider the profile likelihood

$$\mathscr{L}_{\mathscr{P}}(\theta_l \mid y) = \max_{\theta_{j\neq l}} \mathscr{L}(\theta \mid y). \tag{2.20}$$

Thus a $1-\alpha\%$ confidence interval on a single parameter θ_l is:

$$CI_{1-\alpha}(\theta_l \mid y) = \{\theta_l \mid -2(\mathscr{L}_{\mathscr{P}}(\theta_l \mid y) - \mathscr{L}(\hat{\theta} \mid y)) \leq \chi_1^{2^{-1}}(1-\alpha)\}. \tag{2.21}$$

2.5.3 Confidence Bands and Inverse Intervals

The same constraint is used for calculating both inverse intervals and confidence bands. The difference between the two processes is that in the case of prediction intervals the predicted probability z is the parameter; while, the dose s is the parameter of interest for the inverse interval. In other words, the prediction profile likelihood is:

$$\mathscr{L}_{\mathscr{P}}(z \mid s) = \max_{\theta \in \{\theta | \pi(\beta_i, x=s)=z\}} \mathscr{L}(y \mid \theta) \tag{2.22}$$

and the inverse interval profile likelihood is:

$$\mathscr{L}_{\mathscr{P}}(x \mid z) = \max_{\theta \in \{\theta | \pi(\beta_i, x)=z\}} \mathscr{L}(\theta \mid y), \tag{2.23}$$

where $\pi(\beta_i, x)$ is Equation 2.1.

2.5.4 Interval End-Point Location

While the preceding sections described the mathematical definition of the various intervals, they do not suggest an obvious way to find their endpoints. Two variations on a grid search are described here. Although simple, they are not very inefficient (see Chapter 5 for discussion on other alternatives).

To begin, an appropriate step size for the parameter of interest must be selected. Then the model is refit, with the parameter constrained at sequential fixed values in each direction, starting from the maximum penalized estimates. To avoid numerical problems, at each new iteration, the initial values for all parameters in the model can be set to the fitted values from the previous iteration. This is repeated until the profile likelihood corresponding to the proposed parameter value drops below a desired level. Then either the endpoint can be interpolated from the resulting grid of values or step halving can be used to locate the endpoint at the desired precision.

CHAPTER 3

APPLICATION TO REAL DATA

In the previous chapter the algorithms for fitting mixtures of Firth logistic regressions and for calculating confidence and inverse intervals and confidence bands were described. Here, they will be applied to two real datasets. While the datasets are both from dose-response experiments on beetles, the methods extend beyond these limited examples (see Chapter 5 for more discussion).

3.1 *Rhyzopertha dominica*

Rounds [8] characterized the *Rhyzopertha dominica* data (see Table 1.1) as a logistic regression with missing categorical covariates and he fit both a penalized and an unpenalized version. For the sake of completeness and to illustrate that the results are the same, this will be repeated in the context of a mixture of logistic regressions. In addition to the problem of separation, Rounds [8] also noted that the assumption of approximate normality of parameter estimates was probably not met for the penalized model. In this section this will be explored through likelihood profiling and the calculation of profile likelihood-based intervals.

3.1.1 Penalized (Firth) Logistic Regression Mixtures

The unpenalized and penalized mixtures of logistic regression models were applied to the motivating dataset. The parameter estimates controlling the shape of the dose-response curves can be found in Figure 3.1. These results are identical to those of Rounds [8]. The mixing probabilities (α) are similar for the two models. However, the slope (β_0) and intercept (β_1) terms are different for some of the genotypes. For genotype +/B, the estimates for β_0 and β_1 decreased from 89.1 and 30.2 to 11.1 and 3.4, respectively. Similarly, for genotype -/A, the decreases were from 107.8 and 29.8 to 13.5 and 3.3. These differences in parameter estimates translate to dramatic differences in the predicted dose-response curves between the penalized and unpenalized models. In Figure 3.2 the curves corresponding to +/B and -/A genotypes are notably less steep for the penalized fit compared

to the unpenalized fit. The penalized fit does appear to remove the problem of separation that was observed in the traditional fit.

3.1.2 Likelihood Profiling and Profile Likelihood-Based Intervals

One consequence of non-normality of parameter estimates is asymmetry of the log-likelihood around the maximum likelihood, which otherwise would be parabolic. To some degree, the likelihood profile displays the shape of the likelihood function in the neighborhood and in the marginal direction of individual parameters. This will be illustrated for multiple parameters of interest.

3.1.2.1 Confidence Intervals on Model Parameters

First, likelihood profiles of individual parameters were examined. The parameter estimates from the penalized mixture model were exported to Maple and the asymptotic variance-covariance matrix was calculated as the inverse of Equation 2.13. The likelihood of the penalized model was profiled with respect to each of the model parameters, one at a time. Profiling was carried out at equally spaced intervals of 0.05 times the asymptotic standard error. The likelihood was profiled in each direction until it dropped below 4 units less than its maximum.

The profiled log-likelihood of the dose-response curve parameters can be found in Figure 3.3. The profiles for the intercept (β_0) and slope (β_1) parameters are clearly asymmetric for multiple genotypes. Even the mixing proportion (α) is slightly asymmetric for the least frequent genotype, +/A.

As was described in Section 2.5, profile likelihood-based 95% confidence intervals for the parameter estimates were calculated by locating the parameter values where each of the profiles in Figure 3.3 cross a horizontal line at 997.38, which corresponds to $\mathscr{L}(\alpha,\beta) - \chi_1^{2^{-1}}(0.95)/2$. Further, the asymptotic variance-covariance matrix was used to calculate Wald-type intervals for each of the parameters (as from Section 2.4).

Wald-type and profile likelihood-based confidence intervals are compared in Figure 3.4. While all the Wald-type intervals are, by definition, symmetric about their estimates, many parameters have asymmetric profile likelihood-based confidence intervals.

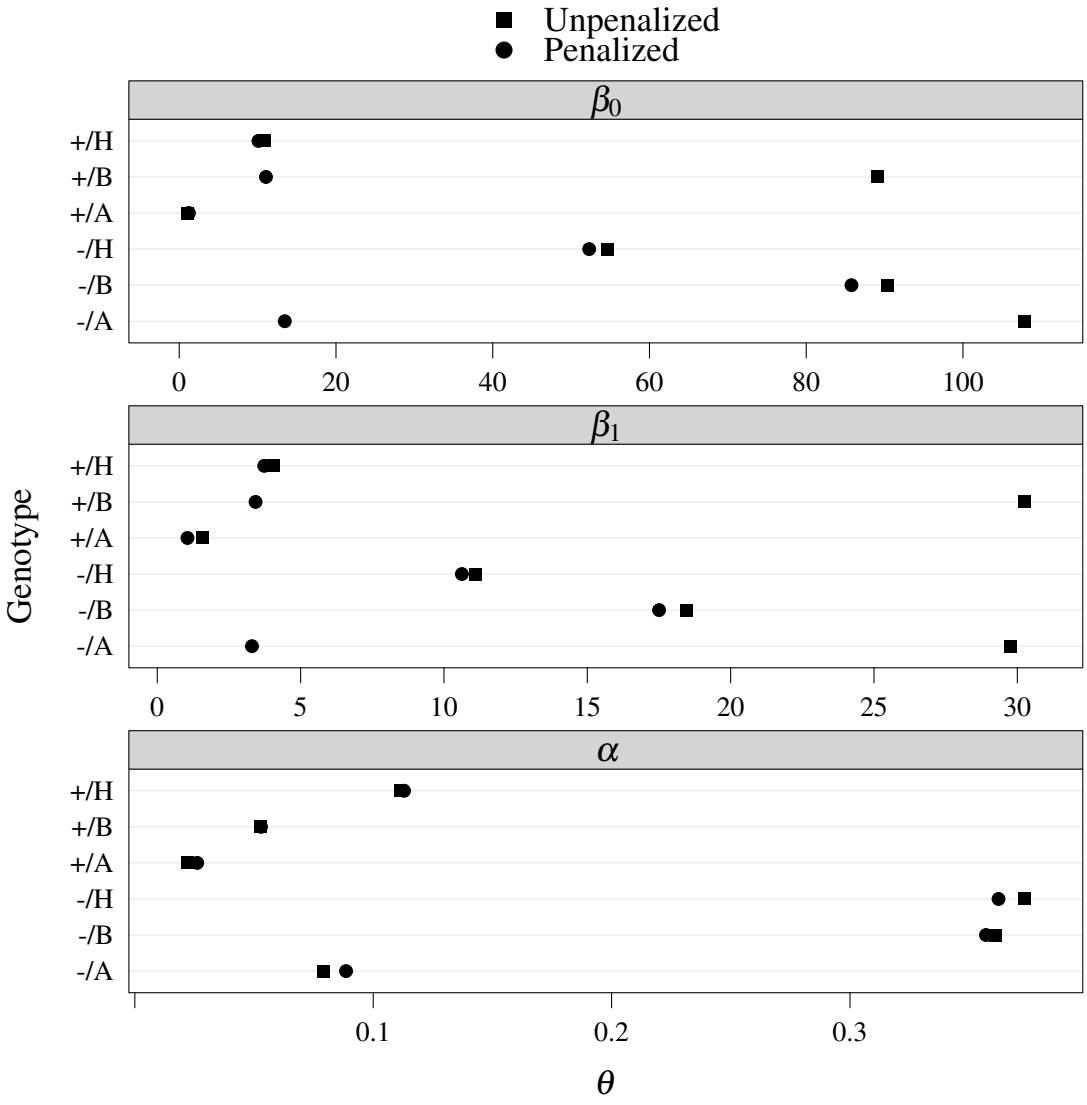

Fig. 3.1: Comparison of the penalized and unpenalized logistic regression mixture parameter estimates for the *Rhyzopertha dominica* data.

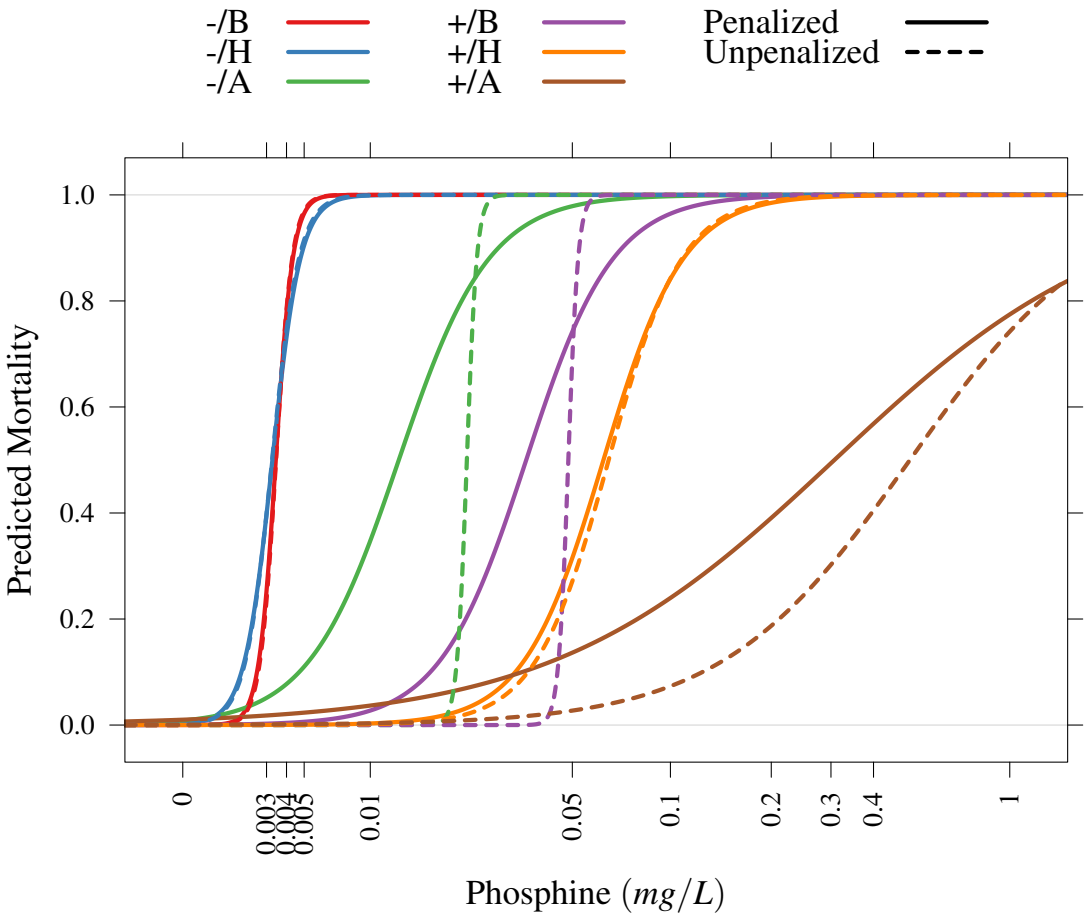

Fig. 3.2: Predicted dose-response curves from the traditional and penalized mixture models for the *Rhyzopertha dominica* data. The solid lines correspond to the penalized fit and the dashed line to the unpenalized. Lines are colored by genotype.

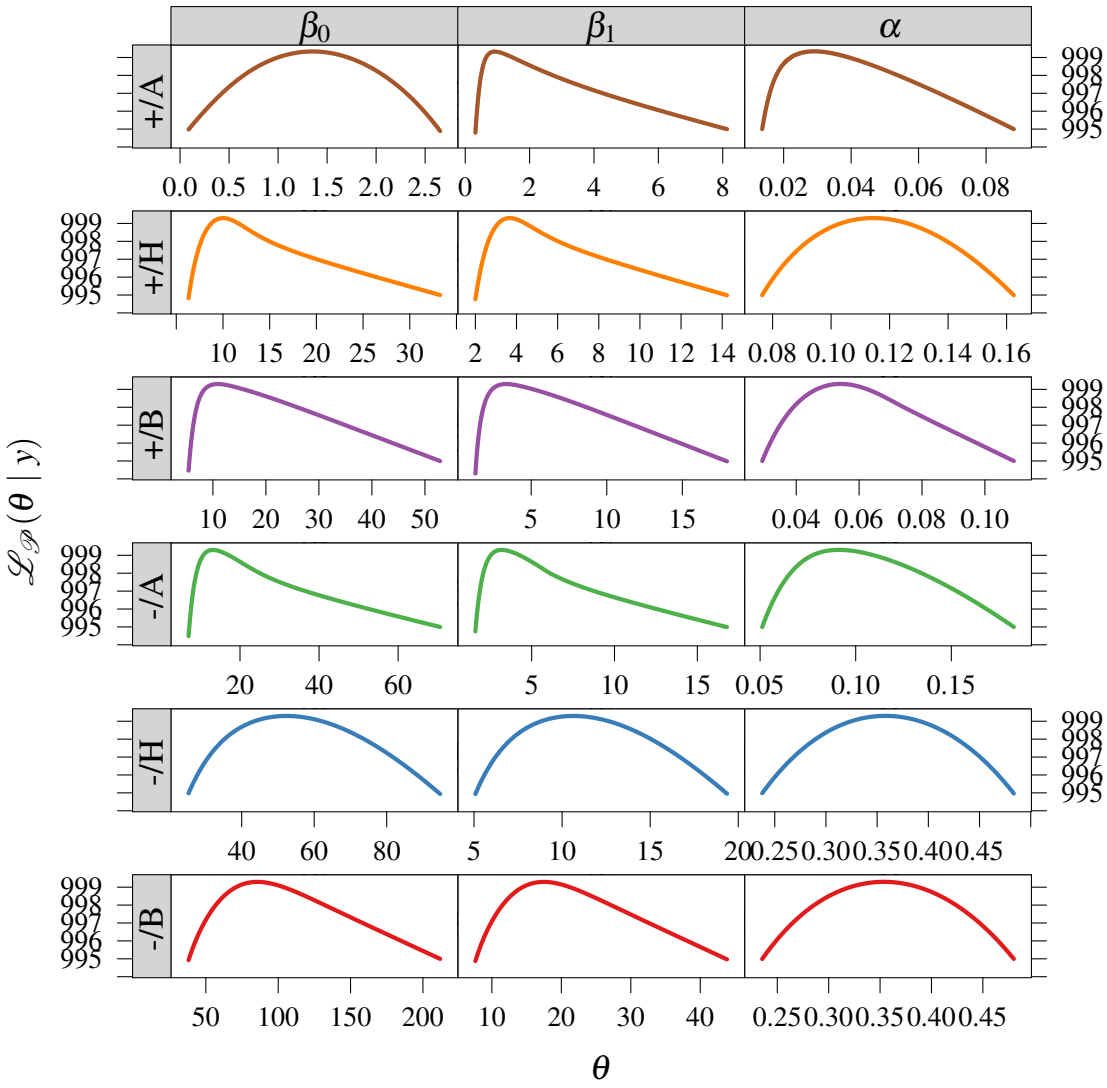

Fig. 3.3: Profiles of the penalized log-likelihood with respect to the logistic regression mixture parameters for the *Rhyzopertha dominica* data.

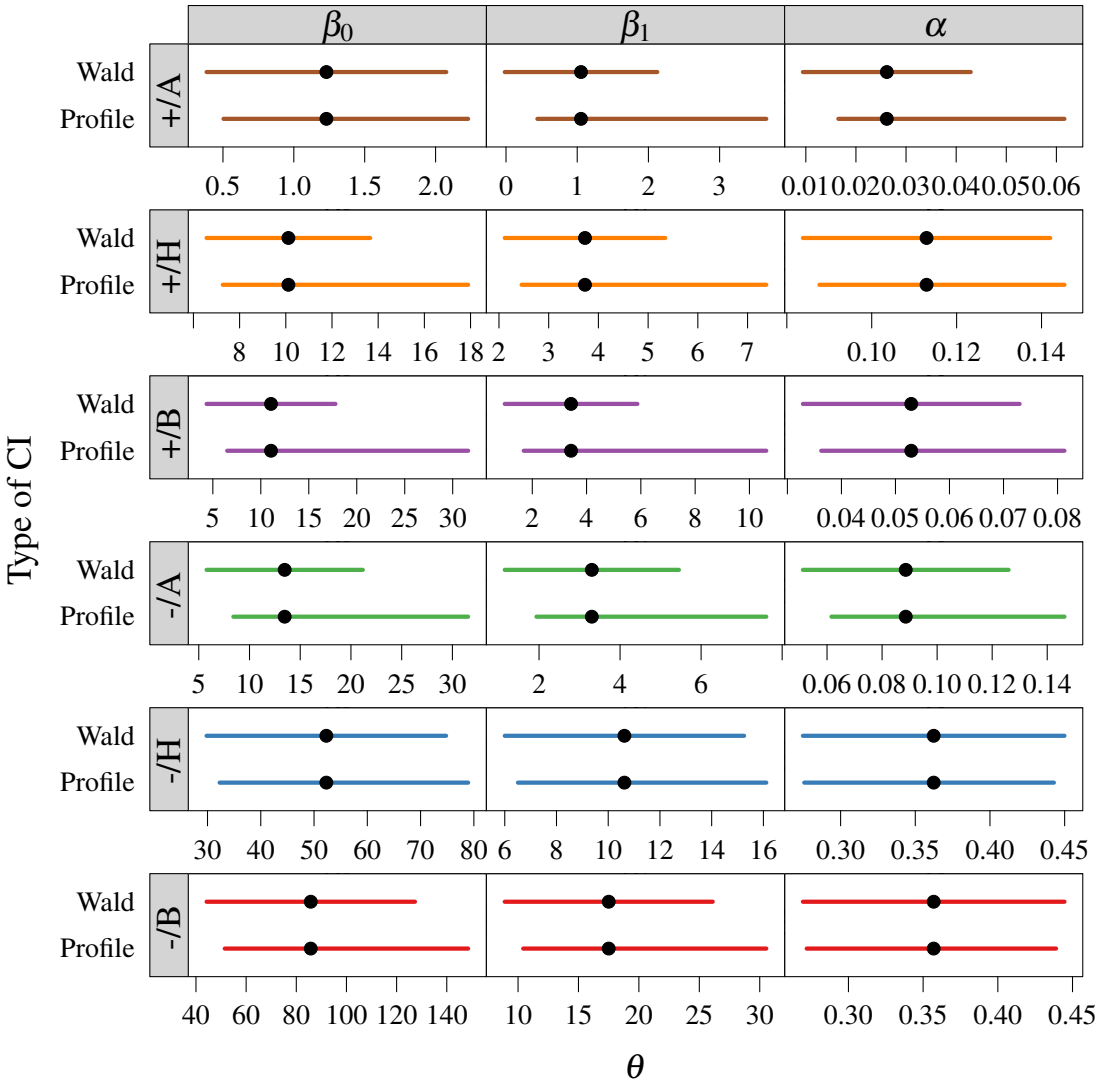

Fig. 3.4: Comparison of Wald-type and profile likelihood-based 95% confidence intervals on parameter estimates from the penalized mixture of logistic regressions for the *Rhyzopertha dominica* data.

3.1.2.2 Inverse Intervals

To further illustrate the utility of likelihood profiling, dose (i.e. inverse probability) was another quantity with respect to which the penalized likelihood from the mixture model was profiled. For each genotype, the dose was constrained to yield predicted mortalities of 0.5, 0.75, 0.9, 0.95, and 0.99. The step size for inverse probability was set to 0.01. The likelihood was profiled in each direction until the penalized likelihood dropped 4 units below its maximum. Inverse intervals were calculated from the profiled likelihoods. So too, were Wald-type and Fieller intervals from the asymptotic variance-covariance matrix.

Likelihood profiles for lethal doses of interest are shown in Figure 3.5. Although not to the degree of the model parameters, these profiles also display some asymmetry. The profile of LD_{95} for genotype +/B is not smooth like the rest of the profiles. This was caused when, during profiling, the likelihood either converged to a local maximum or it failed to fully converge before the algorithm terminated.

Wald-type, Fieller, and profile likelihood-based inverse intervals are shown in Figure 3.6. The slope term was not significant for genotype +/A in the penalized model. As a result, Fieller intervals were not able to be calculated for this genotype. Otherwise, Fieller and profile likelihood-based inverse intervals were similarly asymmetric, with the Fieller intervals being slightly more so in a few cases (e.g. LD_{99} for genotypes -/A and +/B).

3.1.2.3 Confidence Bands

As a final, slightly more complex, example of likelihood profiling, the model was refit with the predicted mortality as the parameter of interest for constraint. For each genotype, the predicted mortality was constrained at values equally spaced on the logit scale from -5 to 5, corresponding to predicted probabilities ranging from approximately 0.0067 to 0.9933. For each constrained predicted mortality value the model was refit, every 0.1 units on the logit scale. Confidence bands were extracted from the profiled likelihood, one predicted value at a time.

Figure 3.7 plots the profiled likelihood as contours across all predicted probabilities and doses. Relative to the predicted mortality, the likelihood is steeper for genotypes -/B and -/H, compared

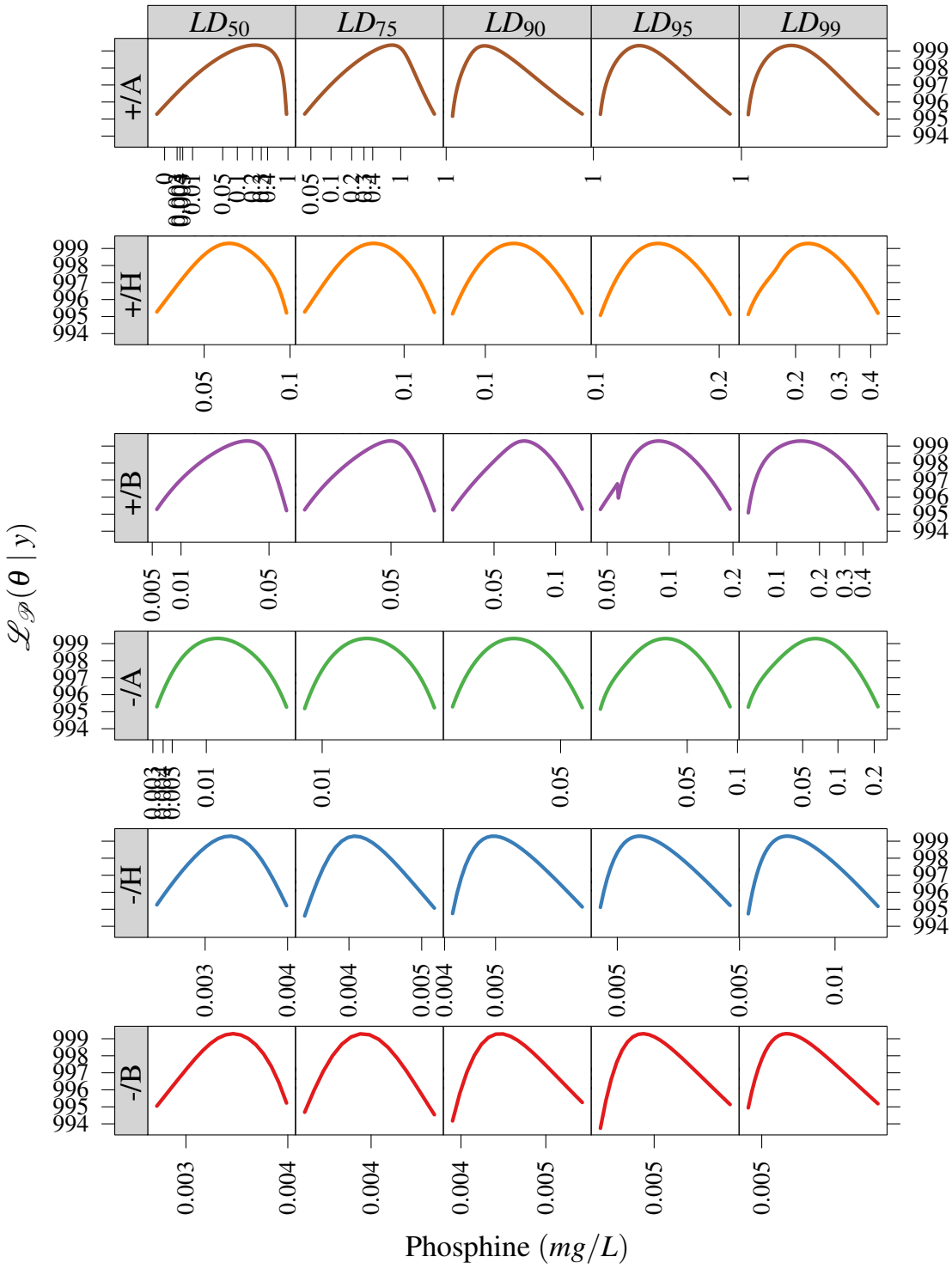

Fig. 3.5: Profiles of the penalized log-likelihood with respect to LD_{50}, LD_{75}, LD_{90}, LD_{95}, and LD_{99} from the logistic regression mixture model for the *Rhyzopertha dominica* data.

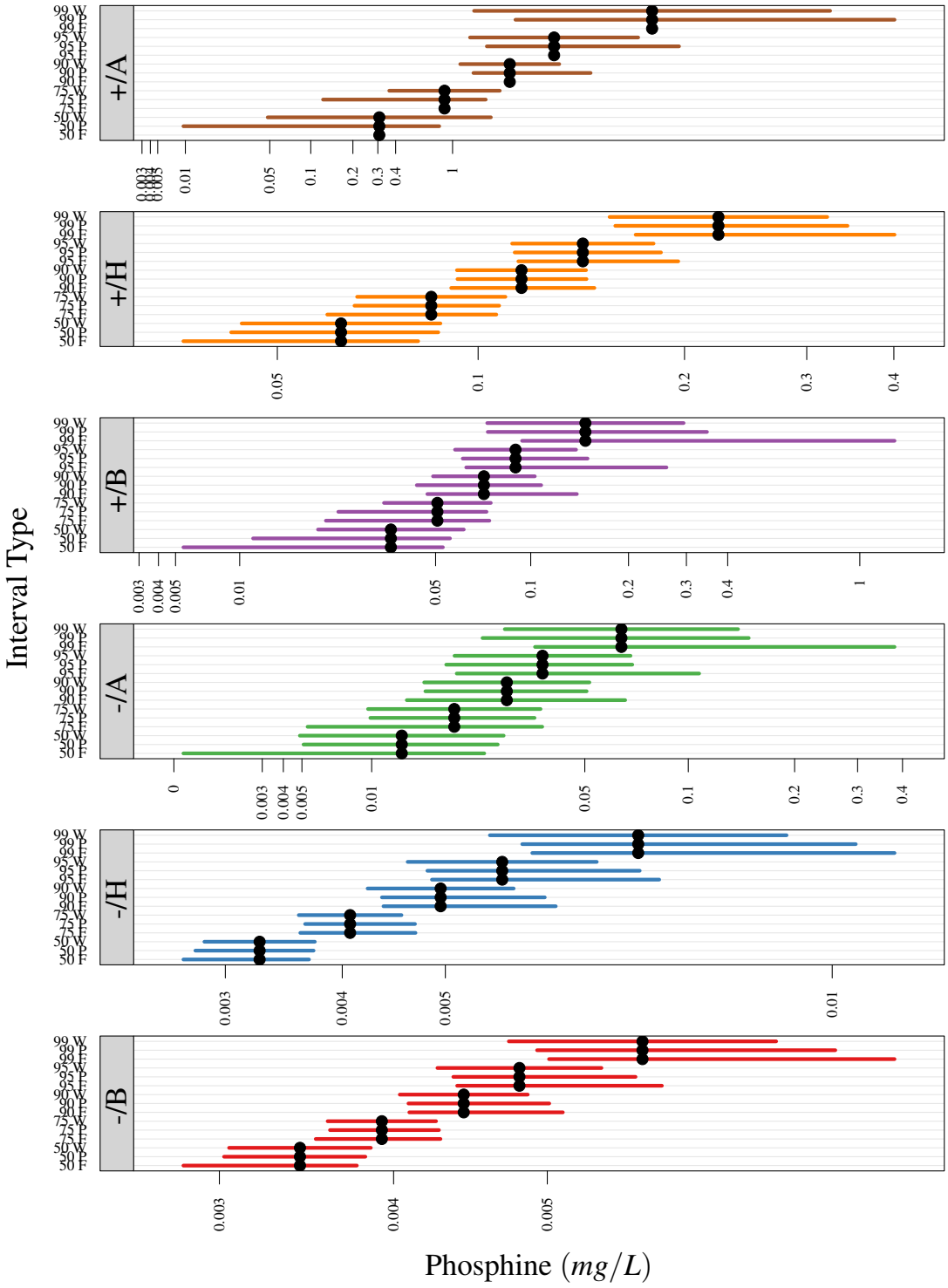

Fig. 3.6: Comparison of Wald-type (W), Fieller (F), and profile likelihood-based (P) inverse intervals for LD_{50}, LD_{75}, LD_{90}, LD_{95}, and LD_{99}. "LD" labeling is suppressed here to avoid excessive redundancy.

to other genotypes. These two genotypes represent the largest proportion of beetles in the dataset (see Table 1.1 and Figure 3.1). The shape of the likelihood represents a combination of the number of beetles at each dose and the degree to which the pattern of mortality fits the logistic regression model.

Wald-type confidence bands were calculated according to Equation 2.15. Both the Wald-type and profile likelihood-based confidence bands are shown in Figure 3.8. The confidence bands are similar except for genotype +/A where the profile likelihood-based bands are more narrow. Very few beetles from this genotype were treated at low doses. The Wald-type bands may not be as flexible as the profile likelihood based bands and therefore do not reflect the changes in the likelihood surface in this area.

3.2 *Tribolium castaneum*

In an experiment similar to that of Schlipalius et al. [1], Jagadeesan et al. [2] identified two genomic locations (tc_*rph1* and tc_*rph2*) linked with phosphine resistance in another grain pest species, *Tribolium castaneum*. They treated 3,910 beetles of the F_4 generation of a dihybrid cross with one of 15 doses of phosphine. In this case, only a random subset of surviving beetles were genotyped. Table 3.1 contains the data for this experiment.

3.2.1 Penalized (Firth) Logistic Regression Mixtures

Mixtures of traditional and penalized logistic regressions were fit to the *Tribolium castaneum* data. Figure 3.9 compares the predicted dose-response curves from the unpenalized and penalized models for each of the genotypes. Even with relatively closely spaced doses the genotypes ss/rr and rr/ss show evidence of separation of points for the unpenalized model. All genotypes have slightly more shallow dose-response curves for the penalized model.

The differences in dose-response curves are perhaps even more clear when the parameter estimates for the models are examined (Figure 3.10). Similar to the *Rhyzopertha dominica* results, the mixing probabilities (α) were comparable; however, the slope and intercept terms for all genotypes are higher for the unpenalized versus the penalized fit. The change is greatest for genotypes ss/rr

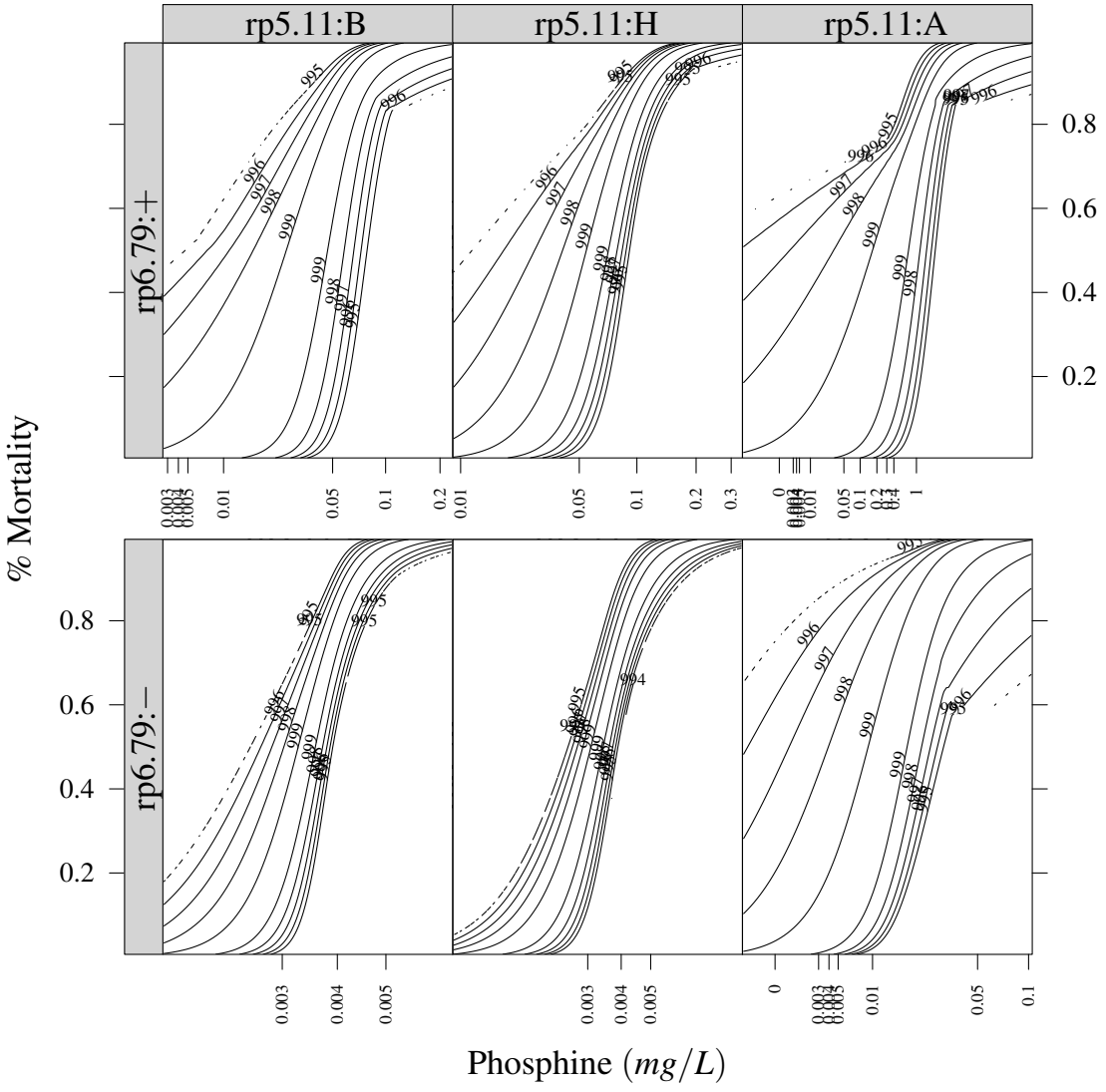

Fig. 3.7: Contour plot of the penalized likelihood, profiled with respect to the predicted mortality for the *Rhyzopertha dominica* data.

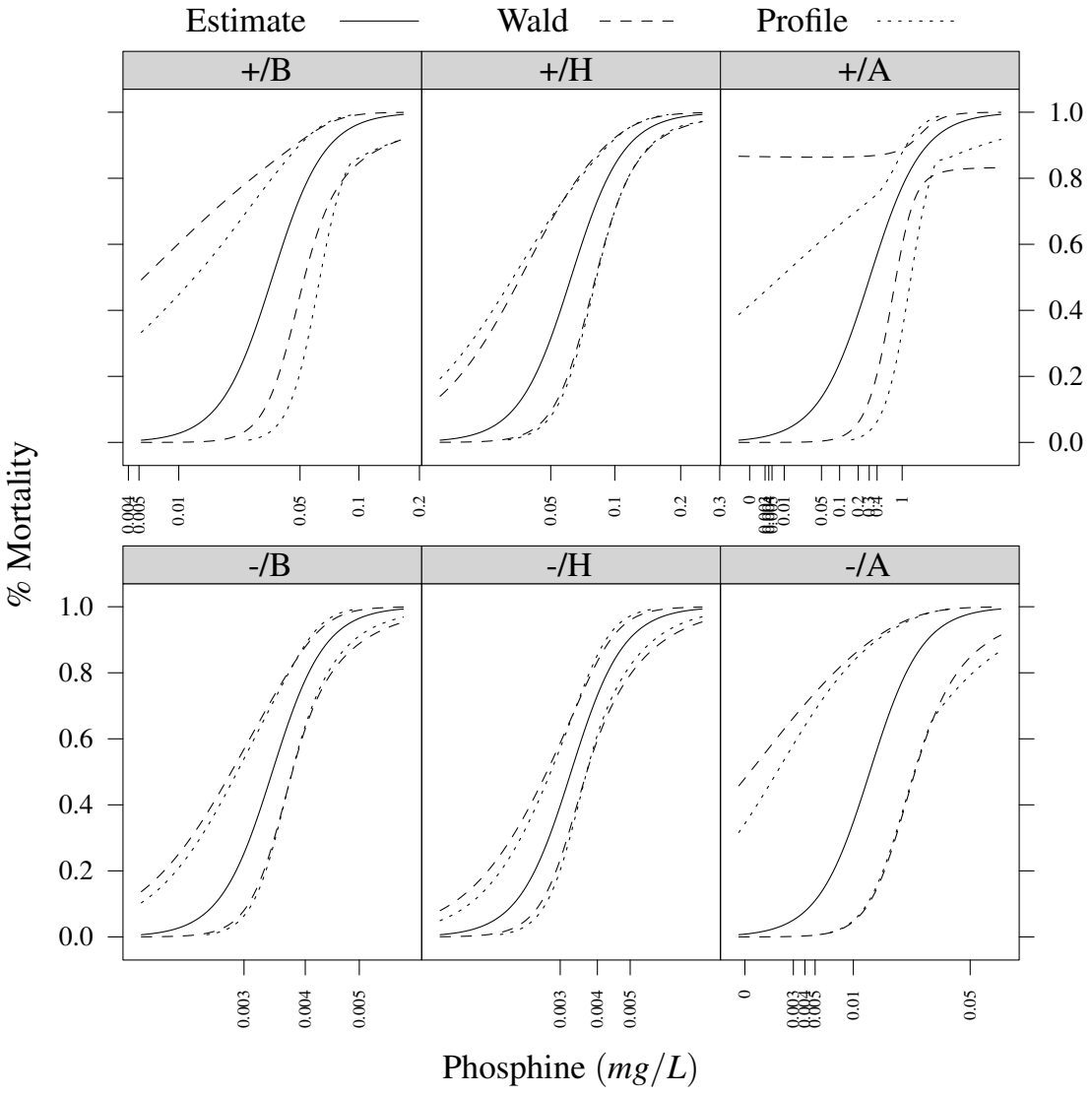

Fig. 3.8: Comparison of Wald-type and profile likelihood-based 95% confidence bands on predicted mortality from the penalized mixture of logistic regression model for the *Rhyzopertha dominica* data.

Table 3.1: Data for a dose-response experiment from Jagadeesan et al. [2] in which *Tribolium castaneum* were treated with phosphine.

Outcome (genotype)		Phosphine Dosage ($\mu g/L$)															Total
		0	8	10	20	30	50	60	100	200	500	800	1000	2000	3000	4000	
Dead	(NA)	0	23	46	111	125	273	174	281	301	289	316	296	306	297	317	3155
Alive	(NA)	109	42	84	22	31	8	7	1	3	1	0	3	1	0	2	314
Alive	(ss/ss)	6	1	0	1	0	0	0	0	0	0	0	0	0	0	0	8
Alive	(sr/ss)	16	11	6	4	0	0	0	0	0	0	0	0	0	0	0	37
Alive	(rr/ss)	15	8	6	15	9	0	0	0	0	0	0	0	0	0	0	53
Alive	(ss/sr)	8	4	2	2	0	0	0	0	0	0	0	0	0	0	0	16
Alive	(sr/sr)	25	21	29	5	1	0	0	0	0	0	0	0	0	0	0	81
Alive	(rr/sr)	18	16	19	31	19	5	12	0	0	0	0	0	0	0	0	120
Alive	(ss/rr)	3	0	2	1	1	7	2	0	0	0	0	0	0	0	0	16
Alive	(sr/rr)	2	3	3	2	10	9	6	8	1	3	0	0	0	0	0	47
Alive	(rr/rr)	0	1	4	7	4	11	3	5	7	7	4	3	3	3	1	63
Total		202	130	201	201	200	313	204	295	312	300	320	302	310	300	320	3910

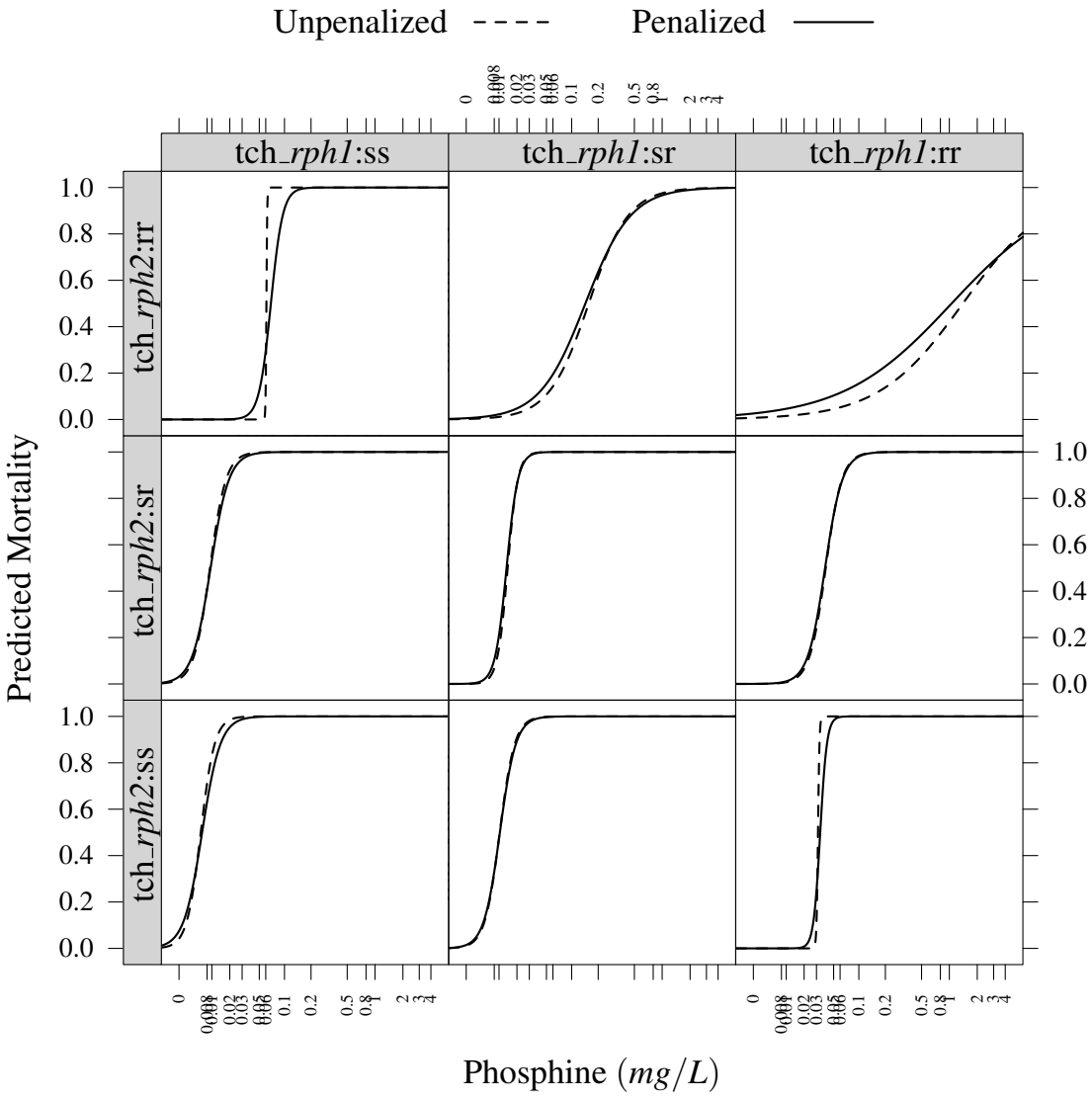

Fig. 3.9: Predicted dose-response curves from traditional and penalized mixture models for the *Tribolium castaneum* data.

and rr/ss. The differences in slope and intercept terms for genotypes ss/rr and rr/ss translate into differences in LD_{99}.

3.2.2 Profile Likelihood-Based Intervals

The parameter estimates from the penalized mixture model were exported to Maple to calculated the asymptotic variance-covariance matrix. However, numeric problems were encountered during the calculation of some of the second partial derivatives of the parameters in Equation 2.13.

As an alternative, the step size for the intercept and slope parameters was based on the asymptotic variance-covariance matrix assuming no missing values and expected genotype frequencies from the penalized fit. The step size for the mixing probability parameters was set to 0.1 on the logit scale. After the first pass of profiling, step halving was used to find the endpoints of the intervals. The same approach as with the *Rhyzopertha dominica* data was employed for step size selection and profiling of inverse and predicted probabilities except that step halving was used to locate the interval endpoints. Additionally, the only inverse intervals that were calculated were for LD_{99}.

Profile likelihood-based confidence intervals on the parameters for the penalized fit on the *Tribolium castaneum* data are shown in Figure 3.10. Likewise, estimates and intervals for LD_{99} are located in the last panel of the same figure. While most of the intervals are basically symmetric, those for the slope and intercept terms for genotypes ss/rr and rr/ss are not. This carries over to the estimates for LD_{99} which are also quite asymmetric.

Profile likelihood-based confidence bands on predicted mortality for each of the genotypes can be found in Figure 3.11. The upper and lower bands for genotype ss/rr are not as smooth as those for the other genotypes. This may be due to the combination of small numbers of labeled beetles, a steep dose-response curve, and only two closely spaced doses in the linear portion of the dose-response curve for this genotype.

3.2.3 Pattern of Inheritance

As it seems reasonable that the penalized model is a better model than the traditional unpenalized model (particularly in the presence of separation of points), it will be used to learn more about the genetic inheritance of resistance to phosphine in *Tribolium castaneum*. If a resistance gene is

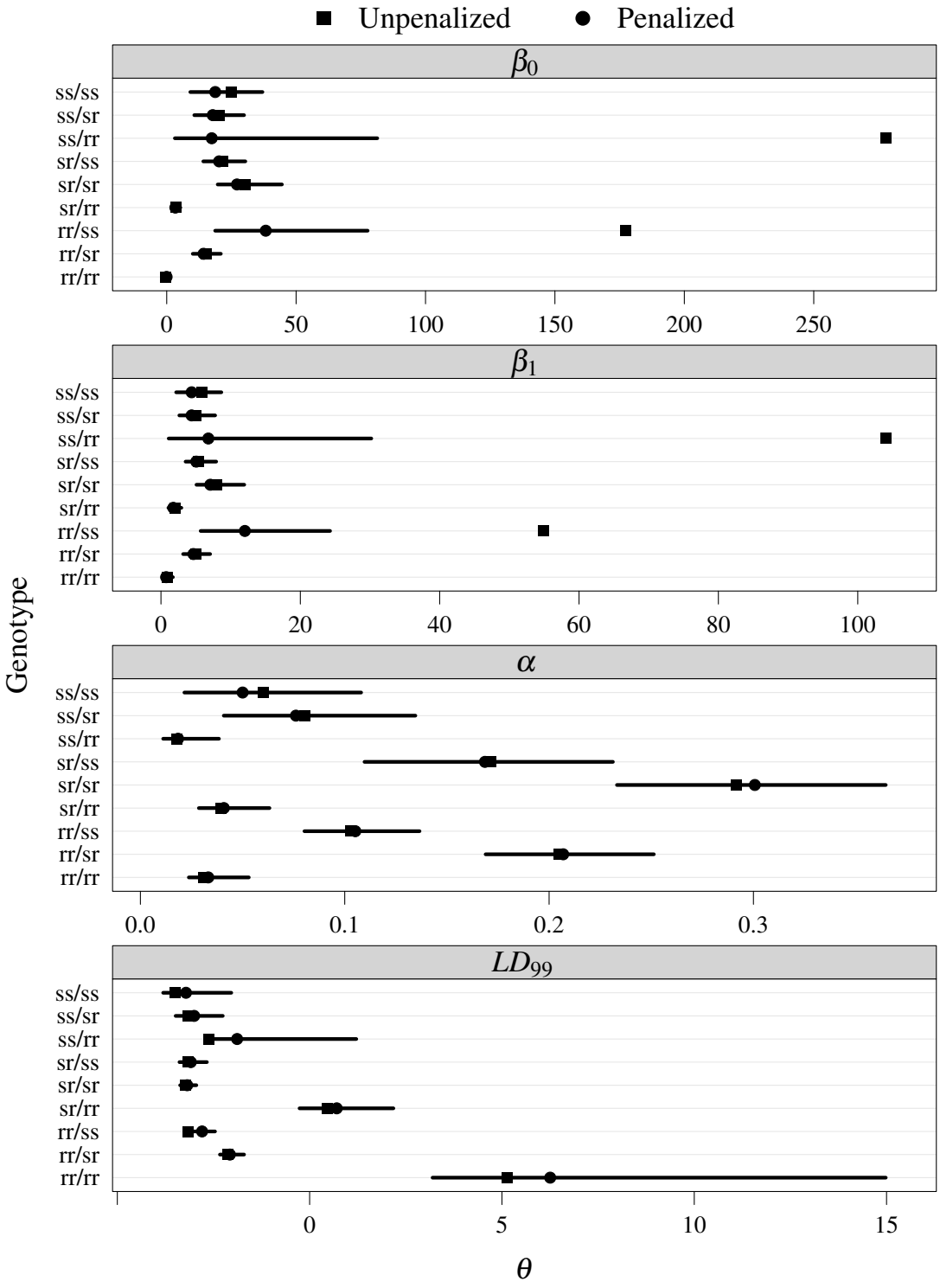

Fig. 3.10: Comparison of the penalized and unpenalized logistic regression mixture parameter estimates for the *Tribolium castaneum* data. The horizontal bars correspond to profile likelihood-based 95% confidence intervals on the parameters from the penalized mixture of logistic regressions.

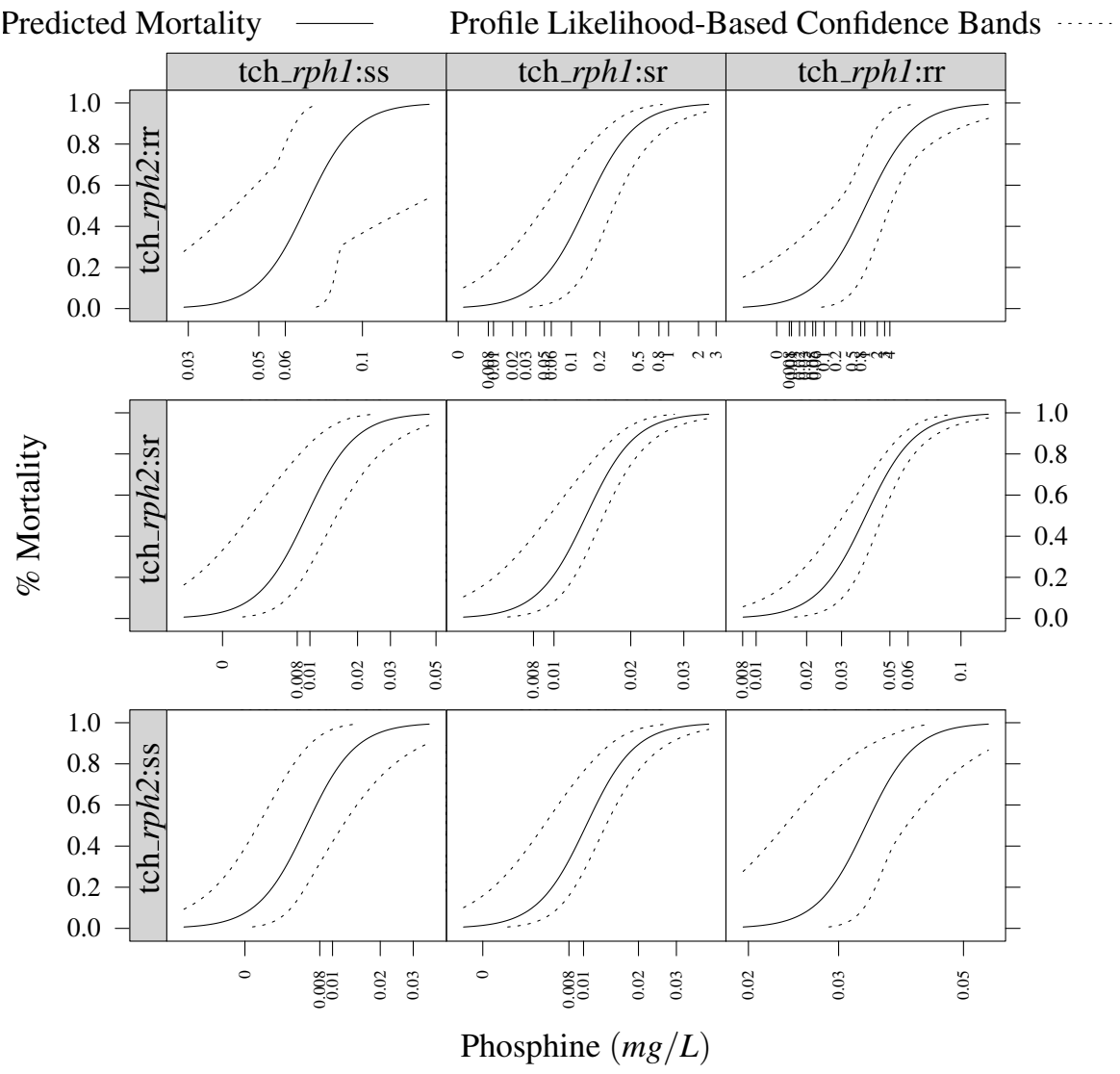

Fig. 3.11: Profile likelihood-based 95% confidence bands on the predicted mortality from the penalized mixture of logistic regressions for the *Tribolium castaneum* data.

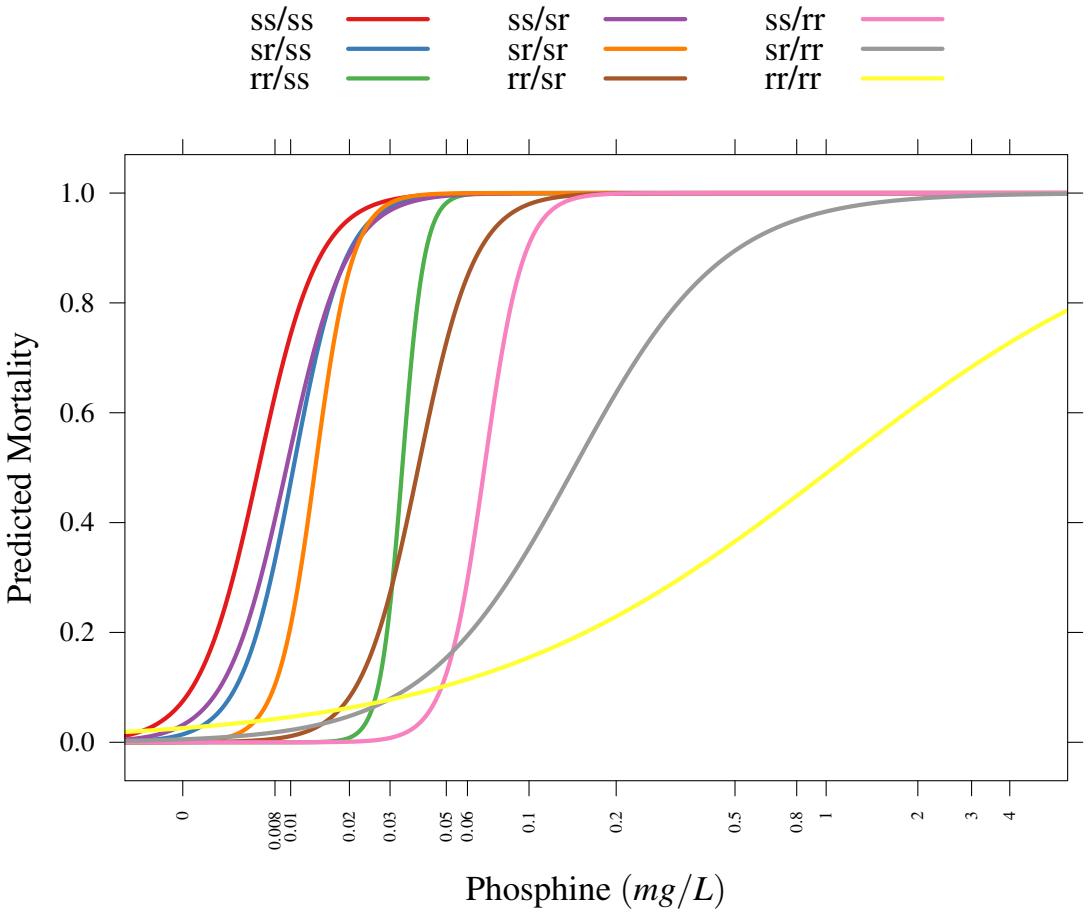

Fig. 3.12: Predicted dose-response curves from the penalized mixture models for the *Tribolium castaneum* data. Lines are colored by genotype.

dominant, only one copy of the allele is needed to confer resistance and the dose-response curves for the heterozygous (sr) and homozygous resistant (rr) genotypes will not be different. Whereas if the resistance gene is recessive, the heterozygous (sr) and homozygous sensitive (ss) genotypes will have indistinguishable dose-response curves. Figure 3.12 reproduces the dose-response curves for the penalized fit. It can be used for a visual assessment as to which of these cases, if either, is correct.

The methodology of Chapter 2 can be used to construct hypothesis tests to help determine the inheritance pattern of phosphine resistance. Constraints can be introduced to force the slope and

intercept parameters to be equal across groups of genotypes.

For each of the two genes, two additional models were fit with linear constraints on β_0 and β_1, thus allowing to carry out the aforementioned hypothesis tests. A penalized likelihood ratio test was carried out for each model compared to the unconstrained model. The tests comparing sr and rr were highly significant for both tc_*rph1* ($\Lambda = 98.5$, df = 6, p-value = 5.3×10^{-19}) and tc_*rph2* ($\Lambda = 286.8$, df = 6, p-value = 5.4×10^{-59}). However, across the genotypes of tc_*rph2*, the dose-response curves corresponding to the homozygous sensitive and heterozygous genotypes of tc_*rph1* were not significantly different ($\Lambda = 8.3$, df = 6, p-value = 0.22). On the other hand, the differences between the dose-response curves corresponding to the homozygous sensitive (ss) and heterozygous genotypes (sr) of tc_*rph2* were small and only moderately significant ($\Lambda = 14.0$, df = 6, p-value = 0.029).

These results suggest that there is insufficient evidence to claim that tc_*rph1* has anything but a recessive pattern of inheritance. While there is much more evidence that the homozygous resistant (rr) genotype of tc_*rph2* is different from the heterozygous (sr) genotype, there is also some evidence that the homozygous sensitive (ss) genotype is different from the heterozygous (sr) genotype. This indicates that tc_*rph2* may have an intermediate mode of inheritance.

3.3 Discussion

A penalized (Firth) logistic regression mixture model was applied to two dose-response experiments on stored grain pests. The penalized model appears to be a better representation of the true phenomenon underlying the data compared to an unpenalized model. In both the *Rhyzopertha dominica* and the *Tribolium castaneum* experiments, the penalized model solved the apparent problem of separation of points, counteracting the irrationally steep behavior of a couple of the dose-response curves, in each case. Of particular note to the applied scientist, the differences between the two models are most extreme in the cases of extreme predicted mortality. One of the main goals for insect pathologists is to estimate the minimum dose necessary to kill virtually all individuals. For these datasets the estimates for LD_{99} were very different for the two models.

Profile likelihood-based intervals were calculated for model parameters, inverse probability

(dose), and predicted mortality. In many cases (although more so in the *Rhyzopertha dominica* data), these intervals were asymmetric, suggesting that the inflexible Wald-type intervals are inadequate. Due to the fact that profiling is more flexible in its representation of the non-parabolic shape of the likelihood near its maximum, these intervals are preferred over the traditional Walt-type intervals.

Computational challenges were encountered while calculating both the Wald-type and profile likelihood-based intervals. In the *Tribolium castaneum* data, some elements of the information matrix were undefined. One of the primary goals of developing the profile likelihood-based intervals methods was to avoid the mathematical analysis necessary for calculating standard errors. Therefore, understanding and solving this problem was not pursued. That is, the availability of the profile likelihood-based confidence intervals provided a useful alternative to the problematic Wald-type intervals. In the *Rhyzopertha dominica* data, there was some evidence of incomplete convergence while profiling dose for one of the genotypes. Smaller step sizes, different initial values, or more restrictive convergence criteria could be applied to prevent this problem.

Applying the method of linear constrained optimization to the penalized EM algorithm not only made profiling possible but it also enabled the execution of some interesting hypothesis tests in the *Tribolium castaneum* data. These hypothesis tests were used to facilitate the understanding of the pattern of heredity of phosphine resistance.

Some factors, within the control of the researchers, were different between the designs of the *Rhyzopertha dominica* and *Tribolium castaneum* dosing experiments. For example, even though the lethality of phosphine is on the same order for the two species, the number and spacing of doses differed. There were more doses that were more uniformly spaced in the *Tribolium castaneum* experiment. Additionally, the number of beetles assigned to each dose was more uniform in the second experiment. While none of the genotypes of the dead beetles were determined, all live beetles were genotyped in the first study and only a random portion in the second. These, and other factors, can be manipulated to improve the experimental design to yield more information without changing the size of the study.

While there is sufficient evidence to argue that the mixture of penalized logistic regressions and profile likelihood-based intervals are better than their traditional counterparts for these two beetle

datasets, it is not obvious in what other circumstances they would be preferred. In Chapter 4 other potential situations will be explored through simulation.

CHAPTER 4

SIMULATION STUDY

In Chapter 3, two mixture models of logistic regressions, unpenalized and penalized, were each fit to data from two similar dose-response experiments. Wald-type and profile likelihood-based confidence intervals were also calculated. Results were compared between the traditional and penalized mixtures and between the Wald and profile confidence intervals. Differences between these pairs of methods were observed. While the proposed methods appeared reasonable, it is impossible to determine, from these datasets alone, which methods are closer to the truth than the traditional methods. Additionally, the strategy used to label subjects in the two experiments was nearly the same and was limited to live beetles.

In this chapter simulation will be used to simultaneously expand the variety of situations where the methods can be applied and determine which methods produce more reliable results. The bias of the two mixture models' point estimates and the coverage probability of the corresponding intervals will be assessed.

4.1 Simulations

A simplified version of the beetle designs from Chapter 3 served as the base scenario for all simulations. The theoretical population of beetles consisted of two genotypes found at equal proportions. One hundred beetles were randomly assigned to each of 10 doses that were equally spaced on the log scale (1:2 serial dilutions starting from $1000\mu g/L$). The total number of beetles in the study was 1000 individuals. The base scenario will be referred to as scenario A.

4.1.1 Number and Spacing of Doses

The base scenario was modified three ways to alter the number and spacing of doses. First, the number of doses was doubled by including doses midway between each of doses used in the base scenario (3:2 serial dilutions starting from $1000\mu g/L$). Each dose was administered to 50 beetles.

This is scenario B. For scenario C, four-fold dilutions were used to yield 5 doses. Two-hundred beetles were treated with each dose in order to keep the total number of beetles the same. The last dosing scenario, scenario D, kept the number of doses at 10, but rather than making them uniform on the log scale the doses were closer together at low doses and further apart at high doses (doses at 0.002, 0.003, 0.004, 0.006, 0.010, 0.016, 0.027, 0.053, 0.125, and 1 $\mu g/L$) .

4.1.2 Allocation of Observations to Doses

Further modifications of the base design changed the allocation of beetles away from uniform. In one scenario more beetles were treated with low doses. Beetles were assigned to doses starting with 190 at the lowest dose and decreasing the number of beetles by 20 at each dose down to 10 beetles at the highest dose (scenario E). The opposite was done to create a scenario where more beetles received the highest doses (scenario F).

4.1.3 Prototypical Genotypes

Genotypes from the *Rhyzopertha dominica* study were used as prototypes for the relationship between phosphine dose and survival. The fitted dose-response curves were used to simulate the number of dead subjects for each dataset. For each scenario described above, five combinations of genotype pairs were simulated: one sensitive and one moderately resistant genotype (-/H and -/A), one sensitive and one highly resistant genotype (-/H and +/A), two moderately resistant genotypes (-/A and +/B), one moderately resistant and one highly resistant genotype (-/A and +/A), and two highly resistant genotypes (+/H and +/A).

4.1.4 Selection of Subjects for Labeling

For every scenario and pair of prototypical genotypes two approaches for labeling the genotypes were employed. First, to be consistent with the two real studies, the genotypes of all dead beetles were censored. Alternatively, only the labels for a random number of beetles at each dose, equal to the number of surviving beetles at the dose, were retained.

4.1.5 Model Fitting and Summarization

For each of the 30 combinations of 6 scenarios and 5 pairs of prototypical genotypes, 1000 datasets were simulated. Both labeling procedures were applied to each dataset. In all, 60,000 partially labeled datasets were produced. The methods from Chapter 1 were applied to each simulated dataset and labeling approach. Mixtures of traditional and penalized logistic regressions were fit to each dataset. The average of each parameter estimate was calculated across replicate datasets. For the penalized model fits, Wald-type and profile likelihood-based confidence intervals were calculated for all the dose-response parameter estimates. Additionally, the LD_{99} and corresponding Fieller, Wald-type, and profile likelihood-based inverse intervals were calculated. The coverage probability of these intervals was determined for each set of replicates.

4.2 Results

Figures 4.1 and 4.2 each show examples of the penalized and unpenalized models fit to 100 datasets selected from the simulated study. Both figures display data simulated from scenario C, where 5 equally spaced doses were each administered to 200 subjects. Additionally, both labeling strategies are shown. In order to visualize all model parameters in a single graph (including the mixing probability), the predicted proportion of total dead beetles is plotted instead of the predicted mortality. In Figure 4.1, the subjects were drawn from a theoretical population consisting of half sensitive (-/H) and half highly resistant (+/A) genotypes. Fitted dose-response curves from two moderately resistant genotypes (-/A and +/B) are shown in Figure 4.2. These two plots provide a reference for illustrating examples of trends found across the 28 additional combinations of scenarios and pairs of prototype samples where 1000 datasets were simulated.

4.2.1 Separation of Points and Bias

The average parameter estimates for all the simulated data are plotted in Figure 4.3.

The slope and intercept terms tend to be biased in the same direction, in effect keeping the inflection point of the dose-response curves centered on the true value across the four combinations of labeling strategies and models. The slope and intercept terms are unbiased for the penalized fit regardless of the labeling strategy (e.g. Figure 4.2) except for in the case of the sensitive genotype

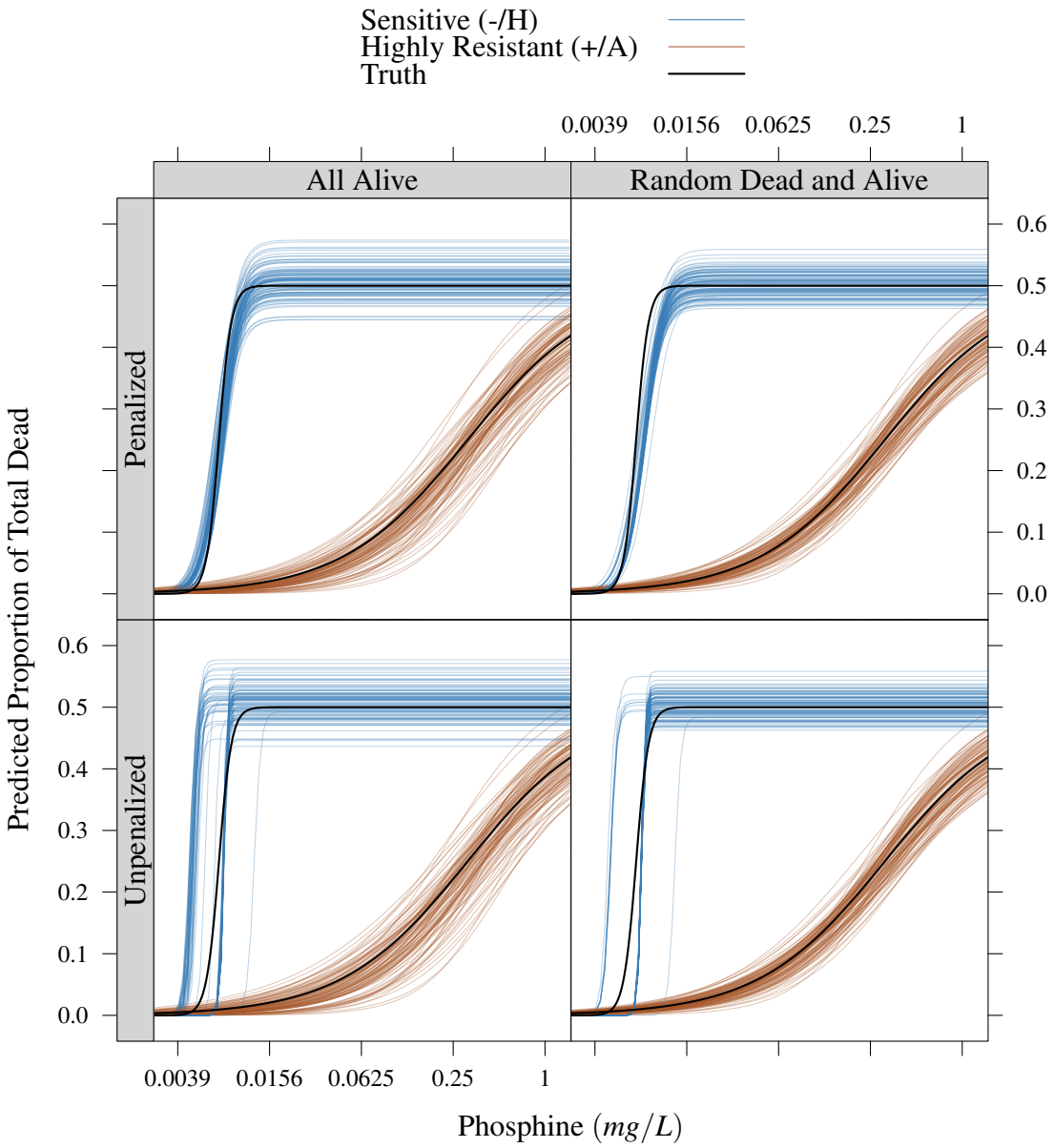

Fig. 4.1: Examples of penalized and unpenalized models fit to 100 simulated datasets with two labeling strategies. The datasets were simulated from scenario C, where five equally spaced doses were each assigned to 200 random subjects from an equally mixed population of sensitive (-/H) and highly resistant (+/A) genotypes.

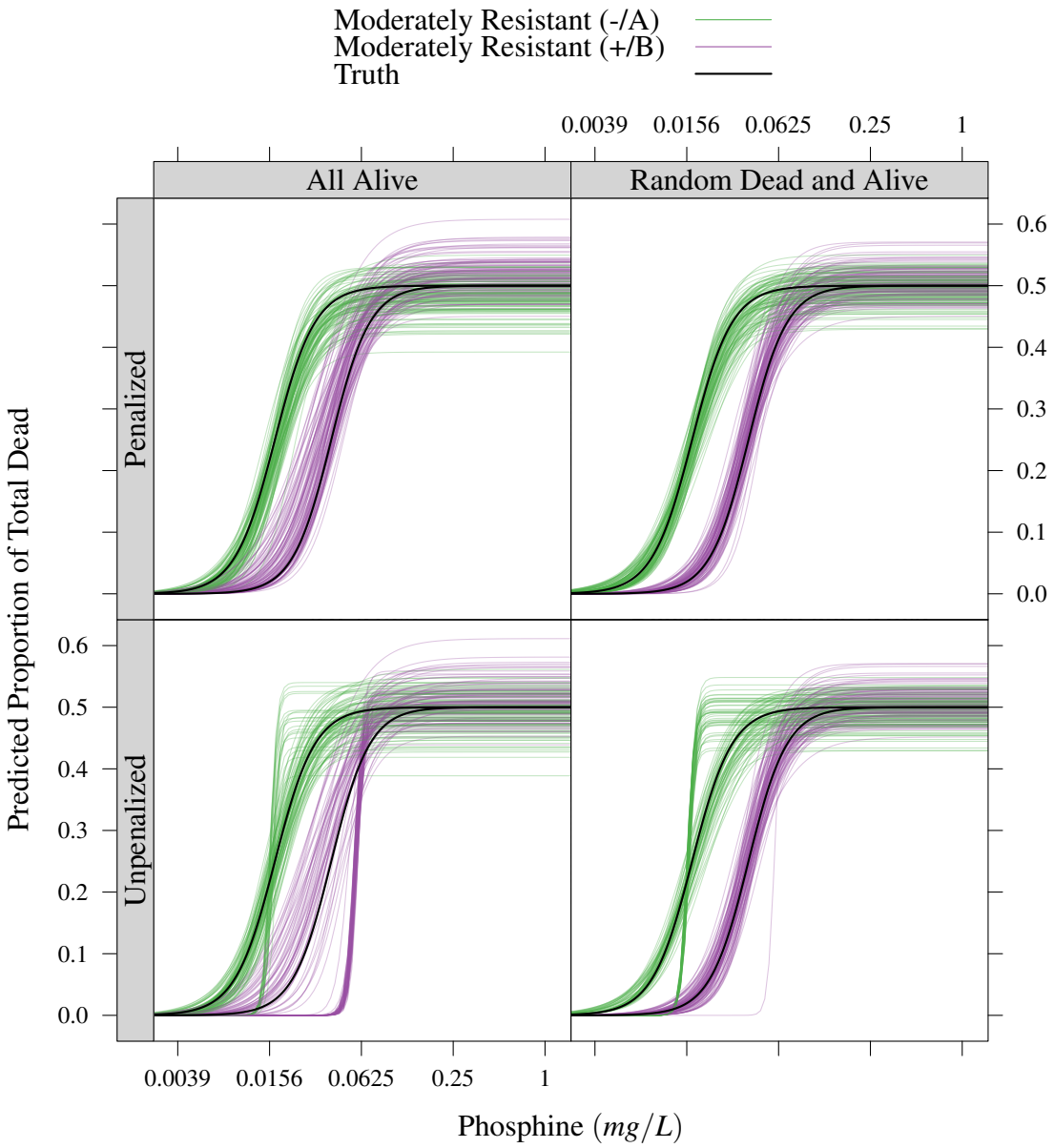

Fig. 4.2: Examples of penalized and unpenalized models fit to 100 simulated datasets with two labeling strategies. The datasets were simulated from scenario C, where five equally spaced doses were each assigned to 200 random subjects from an equally mixed population of two moderately resistant genotypes (-/A and +/B).

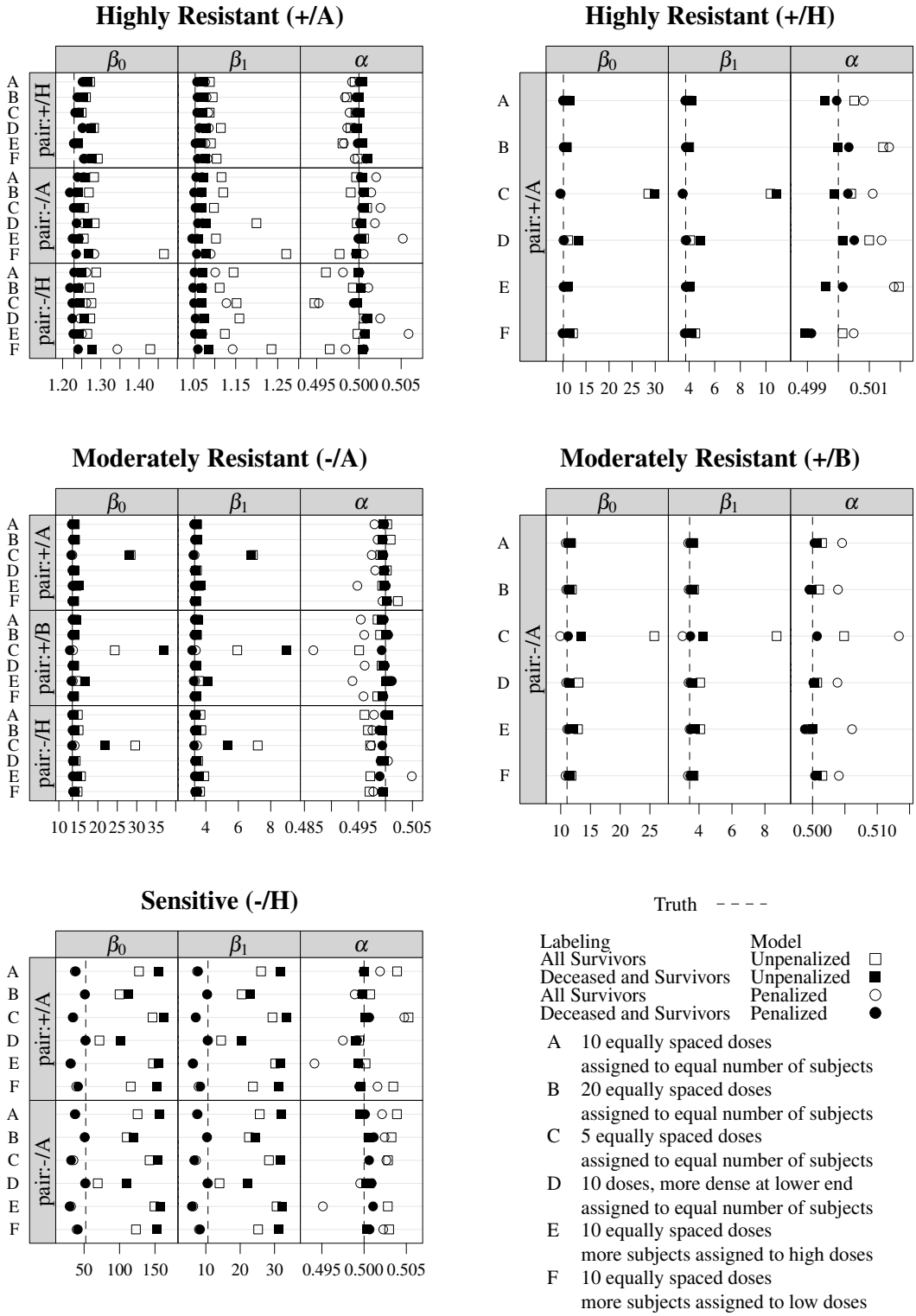

Fig. 4.3: Comparison of penalized and unpenalized parameter estimates averaged across replicates within each simulation scenario (A-F) and labeling strategy.

(-/H, e.g. Figure 4.1). There is an extremely high chance of separation of points for this genotype. The slope from the penalized model is less steep than the truth. Whereas, the unpenalized model yields slope estimates that are much steeper than the truth with a much bigger magnitude than the penalized estimates.

The mixing proportions are virtually unbiased when the alternative labeling strategy is employed. When only live beetles are labeled the mixing proportion can be biased. The penalized model can be worse than the unpenalized model. For instance, genotype +/B in Figure 4.2, is estimated to be more abundant than expected under the "alive only" labeling strategy. The model compensated by finding complete separation of points more often than it should have. With complete labeling, approximately 1.5% of simulated datasets would have quasi-complete separation of points for genotype +/B. However, when only live subjects were labeled, 38.7% of the unpenalized fits showed evidence of separation. Although not as extreme, a similar increase was observed for genotype -/A.

4.2.2 Confidence Intervals and Coverage Probability

There were numeric problems in calculating the asymptotic variance covariance matrix for 6773 of the 30,000 datasets where the "deceased and survivors" labeling strategy was used. These numerical imprecisions were nearly exclusive to, but included almost all, simulations across all scenarios involving the sensitive (-/H) and highly resistant (+/A) genotypes and approximately one third of the simulations involving the sensitive (-/H) and moderately resistant (-/A) genotypes from scenarios B and D.

Figure 4.4 shows the average coverage probability of both the Wald-type and profile likelihood-based confidence intervals for both labeling strategies in every scenario. The coverage probability for Wald-type confidence intervals on the mixing proportions is higher than 95% in virtually every case. On the other hand, the coverage probabilities of the profile likelihood-based confidence intervals for the mixing proportions are very close to 95%. In fact, the profile likelihood-based confidence intervals tend to have better coverage (closer to 95%) than the Wald-type confidence intervals for the slope and intercept parameters as well. The Wald-type confidence intervals have

particularly low coverage in the cases where the penalized estimates were biased, for instance those involving the sensitive genotype (-/H).

The sensitive genotype (-/H) also produces Fieller intervals on LD_{99} with very low coverage probability (see Figure 4.5). Although it is not visible in the figure, neither the Wald-type nor the profile likelihood-based intervals have superior coverage probability of the true LD_{99}, for the other genotypes.

4.3 Discussion

A simulation study was carried out to evaluate the performance of penalized (Firth) logistic regression mixtures and profile likelihood based-confidence intervals over more traditional methods in the context of dose-response experiments.

Separation of points and bias in the slope (and intercept) terms of dose-response curves occurs when too few doses are applied to too few subjects in regions where the dose-response curve is expected to be steep. The penalized model reduces or eliminates this problem. When only surviving subjects are labeled the mixing proportion can also be biased. The penalized model does not appear to fix this bias; however, the alternative labeling strategy (a random sampling of surviving and deceased subjects) does.

Similar to the *Tribolium castaneum* analysis, problems with numerical precision were encountered when calculating the asymptotic variance-covariance matrix. While such calculations are necessary for calculating both Wald-type and Fieller inverse intervals, profile likelihood-based intervals avoid them and were able to be calculated. Additionally, the Wald-type and Fieller intervals were more susceptible to small biases that remained in the penalized fit in extreme cases. Overall, the profile likelihood-based intervals had better coverage probability than the traditional intervals.

While not fully examined here, there is some evidence that the width and symmetry of intervals may be improved using the alternative labeling strategy (Figure 4.2). Additional simulation studies and applications to real data should be used to explore the effect of other labeling strategies, experimental designs, and sample size limitations on penalized logistic regression mixtures and corresponding profile likelihood based intervals.

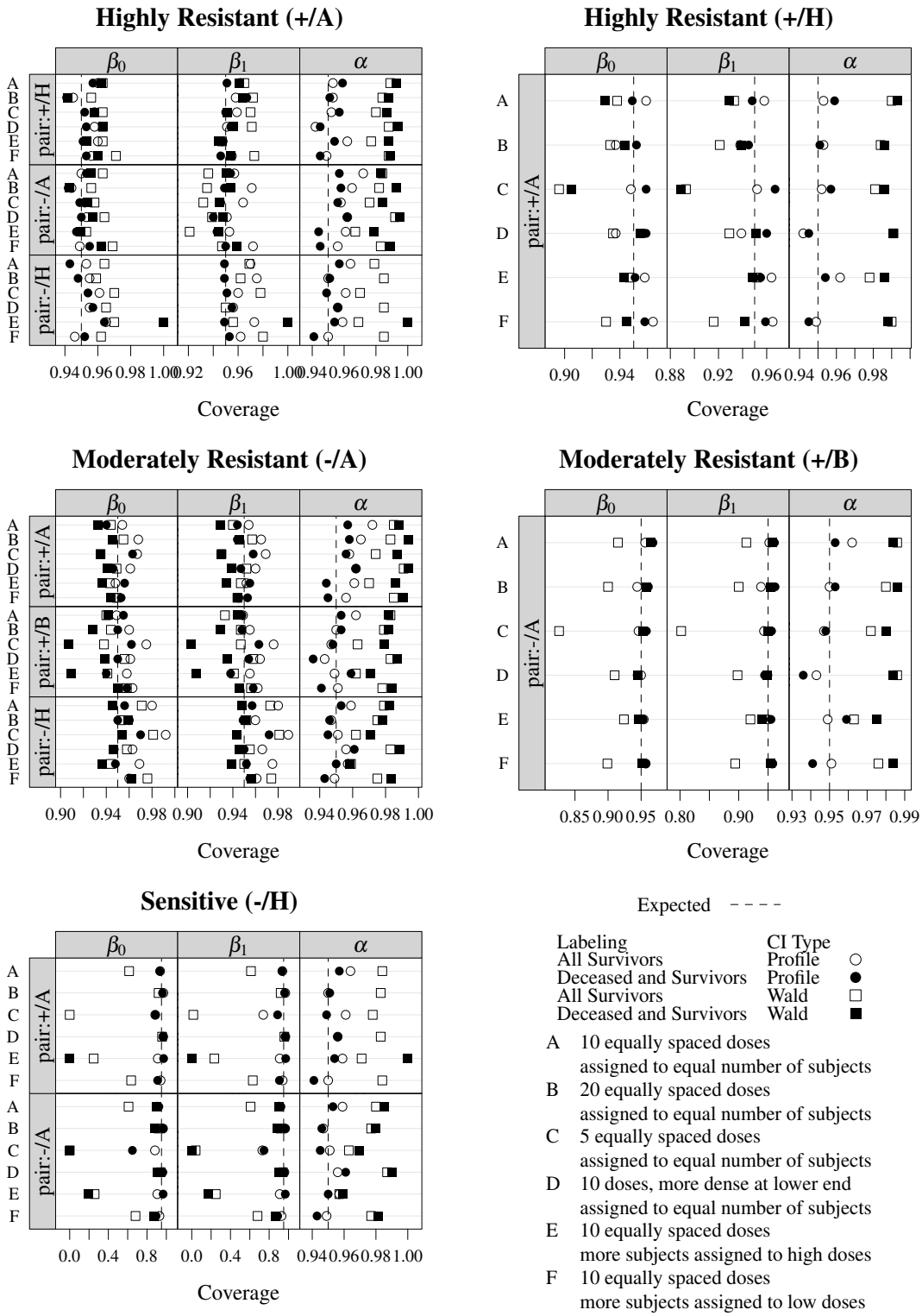

Fig. 4.4: Comparison of average coverage of Wald-type and profile likelihood-based confidence intervals for all simulation scenarios (A-F) and labeling strategies.

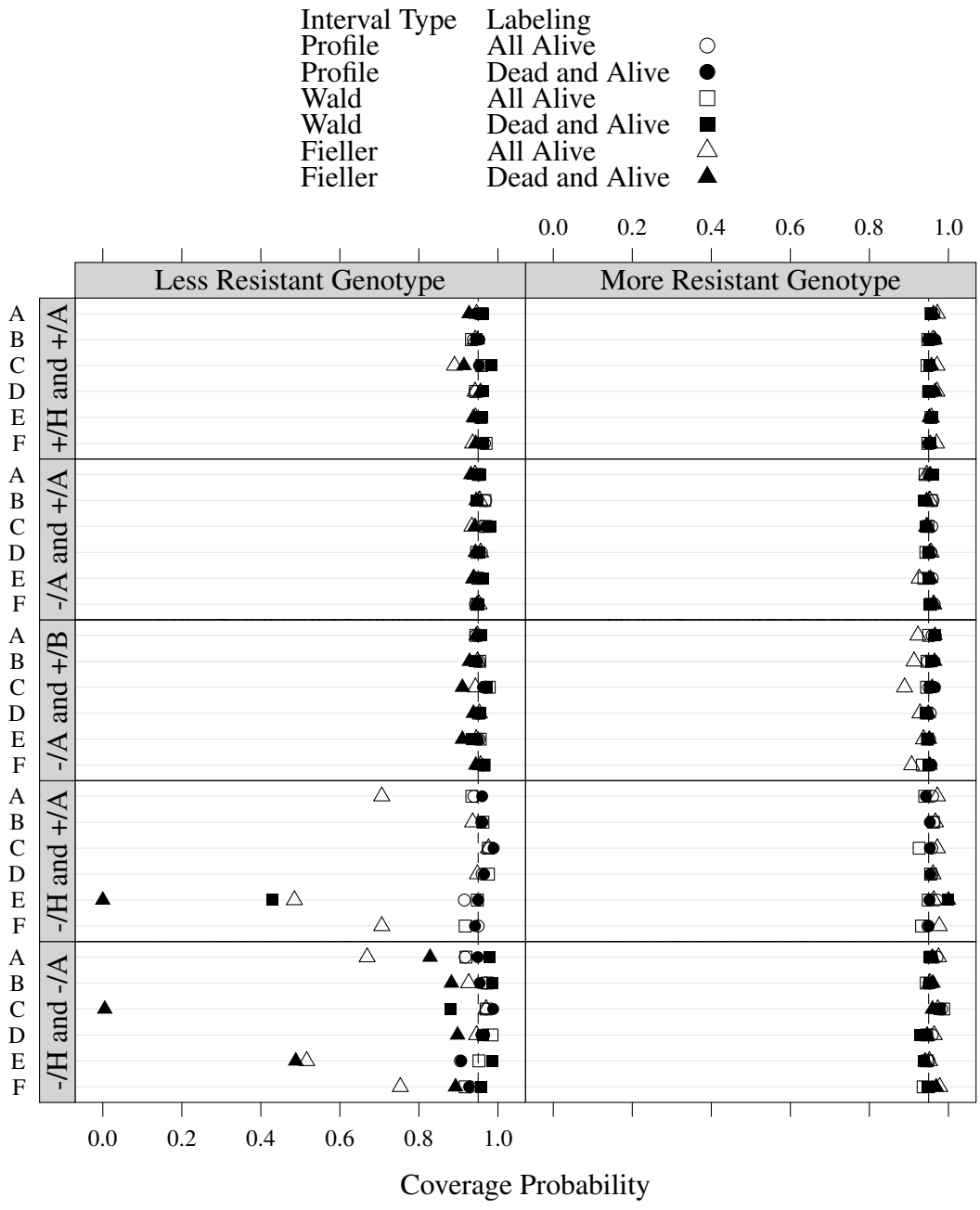

Fig. 4.5: Comparison of average coverage probability of Fieller, Wald-type, and profile likelihood-based inverse intervals on LD_{99} for all simulation scenarios and labeling strategies.

CHAPTER 5

DISCUSSION

The *Rhyzopertha dominica* fumigation experiment provides an interesting dataset where a novel combination of statistical methods can be applied. Perhaps the most obvious characteristic of the data is that the genotypes of the beetles that died were unobserved. Additionally, the number of events are not only extremely low and high, relative to the non-events for some doses, but there is also evidence of complete or quasi-complete separation of the data. Low effective sample size and parameter estimates near their boundaries also call into question the use of Wald-type confidence intervals. The novelty of this dissertation is the simultaneous employment of the EM algorithm, penalized (Firth) logistic regression, and likelihood profiling to overcome the challenge of missing data, to ensure unique and consistent parameter estimates, and to obtain confidence intervals with suitable coverage probability.

Previous to this dissertation, the beetle data had been characterized as a penalized logistic regression model with missing categorical covariates. Here, a more natural characterization of a mixture of penalized logistic regression models with partial labeling of the mixture components was used. Both the mixture of penalized logistic regression models and the partial labeling of logistic regression mixture components are novel applications. The benefits of the penalty were illustrated in the presence of missing values using real data as well as simulations. For the first time, profile penalized likelihood-based confidence intervals were calculated in the presence of missing values. Additionally, profile likelihood based confidence bands and inverse intervals had never been calculated in the mixture setting with or without the Firth penalty. These intervals were shown to behave as expected in real data and in a simulation study.

5.1 Benefits of the Proposed Methods

It was illustrated that mixed penalized (Firth) regressions overcome the previously observed problem of separation of points. Implementation of the penalized model is barely more complex

than the unpenalized model. In extreme cases of separation, the penalized model yielded slope and intercept estimates that were biased in the opposite direction of the unpenalized model; although, the magnitude of the bias was much less extreme.

In the setting of mixed logistic regressions, profile likelihood-based intervals have advantages over both Wald-type and Fieller intervals. Advanced analytical or computational expertise or a combination of the two are required for the calculation of the asymptotic variance-covariance matrix, on which both traditional intervals rely. Profile likelihood-based intervals have consistently better coverage probability than Wald-type intervals and they are less susceptible to bias in the underlying parameters. Additionally, the same method of linearly constrained optimization can be used to calculate various kinds of intervals, while understanding of the sampling distribution of the parameter of interest and an application of the delta method is required to generate appropriate Wald-type intervals. Finally, likelihood profiling serves two useful purposes, in that it can be used to provide a useful diagnostic for the appropriateness of symmetric Wald-type intervals and, at the same time, to calculate better confidence intervals when they are not.

5.2 Limitations of the Proposed Methods

The major drawback of the methods introduced in this dissertation is computational expense. Maximizing individual models is not particularly time consuming; however, profiling requires multiple maximizations of the same likelihood with different constraints. One must be able to move along the likelihood surface efficiently. If aggressive step sizes are taken and reasonable initial values are not proposed, the complete data information matrix in the "M" step becomes numerically singular. To avoid this problem in this study, small steps were always taken, thereby increasing the time needed for computation. For example, on a 2.5 GHz with 6 GB of RAM, it required approximately four and a half days of constant computation to fit all simulated datasets for a single labeling strategy and to profile the likelihoods to calculate confidence intervals of the intercepts, slopes, and mixing proportions of the same.

Methods for speeding up the EM algorithm and for the efficient localization of the endpoints of profile likelihood-based confidence intervals have been proposed. These usually involve knowledge

of the observed data information matrix. For example, Venzon and Moolgavkar [28] proposed a method that maximizes the observed likelihood directly using a Newton-Raphson method and the information matrix from Oakes [7], to obtain profile likelihood-based confidence interval endpoints. Similar methods could be used to replace the EM algorithm entirely for both model fitting and endpoint location.

5.3 Alternatives to the Proposed Methods

This dissertation presents an analysis of mixtures of penalized (Firth) logistic regression models from the frequentist perspective. An alternative would be to fit this model within the Bayesian framework. The Firth penalty is the same as imposing Jeffreys prior on the parameters in the context of a Bayesian analysis [10]. Ibrahim and Laud [29] described the analysis of generalized linear models using Jeffreys prior. Chen et al. [30] provided analytical and computational details for logistic regression with Jeffreys prior. From the Bayesian perspective, missing data are treated like additional parameters. Models with missing data are fit by data augmentation. Future work could include comparing the results for the frequentist and Bayesian model fits.

5.4 Extensions of the Proposed Methods

While this dissertation only employs application to data from dose-response experiments of phosphine on beetles, mixtures of penalized (Firth) logistic regression models could also be more widely applied. In a vignette for their R package flexmix, Grün and Leisch [31] provide four example datasets that have been analyzed using mixtures of logistic regressions. (One of which was an experiment involving *Tribolium castaneum*, although not a dose-response experiment. Another was a dose-response experiment, but it did not involve stored grain pests.) It would be interesting to see if any conclusions would change as a result of applying the proposed methods.

It is also conceivable to use the proposed methods in studies investigating genome-wide association of dose-response such as discovery of potential drug, insecticide, or pesticide targets; or for finding biomarkers of resistance to the same. While all subjects in a representative set selected from a population would need to be randomly assigned to a dose of the agent of interest, only a subset of subjects would need to be genotyped for the common genetic variants distributed across the genome.

The methods developed here could be used to identify the markers with the strongest association with response while saving on the cost of evaluating all subjects. Additional investigation is needed to determine if such a design would have any advantages over current practice.

5.5 Conclusion

Mixtures of penalized logistic regressions show very little bias in comparison to mixtures of traditional logistic regression even in very extreme study situations. Profile likelihood-based intervals are more flexible, easier to implement, and have better coverage than asymptotic variance-covariance based intervals. Therefore, it is suggested that mixed penalized (Firth) logistic regressions and profile likelihood-based intervals be used in place of mixed logistic regressions and Wald-type intervals, particularly when there is the potential for separation of points. These methods are implemented in R code provided in Appendix A, including a demonstration using the actual *Rhyzopertha dominica* data in Appendix A.5.

REFERENCES

[1] Schlipalius DI, Cheng Q, Reilly PEB, Collins PJ, and Ebert PR (2002). Genetic linkage analysis of the lesser grain borer *Rhyzopertha dominica* identifies two loci that confer high-level resistance to the fumigant phosphine. Genetics 161(2): 773–782.

[2] Jagadeesan R, Fotheringham A, Ebert PR, and Schlipalius DI (2013). Rapid genome wide mapping of phosphine resistance loci by a simple regional averaging analysis in the red flour beetle, *Tribolium castaneum*. BMC Genomics 14(650).

[3] Berkson J (1944). Application of logistic function to bio-assay. Journal of the American Statistical Association 39(227): 357–365.

[4] Dempster AP, Laird NM, and Rubin DB (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological) 39(1): 1–38.

[5] Ibrahim JG (1990). Incomplete data in generalized linear models. Journal of the American Statistical Association 85(411): 765–769.

[6] Stevens JR and Schlipalius DI (2006). Dose-response modeling with marginal information on a missing categorical covariate. In Proceedings of Conference on Applied Statistics in Agriculture, pp. 18–32. Kansas State University.

[7] Oakes D (1999). Direct calculation of the information matrix via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological) 61(2): 479–482.

[8] Rounds J (2009). Logistic models with missing categorical covariates. Utah State University, Logan, Utah. Unpublished Master's Report.

[9] Albert A and Anderson JA (1984). On the existence of maximum likelihood estimates in logistic regression models. Biometrika 71(1): 1–10.

[10] Firth D (1993). Bias reduction of maximum likelihood estimates. Biometrika 80(1): 27–38.

[11] Heinze G and Schemper M (2002). A solution to the problem of separation in logistic regression. Statistics in Medicine 21(16): 2409–2419.

[12] Green PJ (1990). On use of the EM for penalized likelihood estimation. Journal of the Royal Statistical Society. Series B (Methodological) 52(3): 443–452.

[13] Wang P and Puterman ML (1998). Mixed logistic regression models. Journal of Agricultural, Biological, and Environmental Statistics 3(2): 175–200.

[14] Redner RA and Walker HF (1984). Mixture densities, maximum likelihood and the EM algorithm. SIAM Review 26(2): 495–239.

[15] Wald A (1942). Asymptotically shortest confidence intervals. The Annals of Mathematical Statistics 13(2): 127–137.

[16] Fieller EC (1954). Some problems in interval estimation. Journal of the Royal Statistical Society. Series B (Methodological) 16(2): 175–185.

[17] Hudson DJ (1971). Interval estimation from the likelihood function. Journal of the Royal Statistical Society. Series B (Methodological) 33(2): 256–262.

[18] Kalbfleish JD and Sprott DA (1970). Applications of likelihood methods to models involving large numbers of parameters. Journal of the Royal Statistical Society. Series B (Methodological) 32(2): 175–208.

[19] Patefield WM (1977). On the maximized likelihood function. Sankhyā: The Indian Journal of Statistics, Series B (1960-2002) 39(1): 92–96.

[20] Beale EML (1960). Confidence regions in non-linear estimation. Journal of the Royal Statistical Society. Series B (Methodological) 22(1): 41–88.

[21] Neale MC and Miller MB (1997). The use of likelihood-based confidence intervals in genetic models. Behavior Genetics 27(2): 113–120.

[22] Meeker WQ and Escobar LA (1995). Teaching about approximate confidence regions based on maximum likelihood estimation. The American Statistician 49(1): 44–53.

[23] Bjørnstad JF (1990). Predictive likelihood: A review. Statistical Science 5(1): 242–265.

[24] Kreutz C, Raue A, and Timmer J (2012). Likelihood based observability analysis and confidence intervals for predictions of dynamic models. BMC Systems Biology 6(120).

[25] Williams DA (1986). Interval estimation of the median lethal dose. Biometrics 42(3): 641–645.

[26] Alho JM and Valtonen E (1995). Interval estimation of inverse dose-response. Biometrics 51(2): 491–501.

[27] Kim DK and Taylor JMG (1995). The restricted EM algorithm for maximum likelihood estimation under linear restrictions on the parameters. Journal of the American Statistical Association 90(430): 708–716.

[28] Venzon DJ and Moolgavkar SH (1988). A method for computing profile-likelihood-based confidence intervals. Journal of the Royal Statistical Society. Series C (Applied Statistics) 37(1): 87–94.

[29] Ibrahim JG and Laud PW (1991). On Bayesian analysis of generalized linear models using Jeffreys's prior. Journal of the American Statistical Association 86(416): 981–986.

[30] Chen MH, Ibrahim JG, and Kim S (2008). Properties and implementation of Jeffreys's prior in binomial regression models. Journal of the American Statistical Association 103(484): 1659–1664.

[31] Grün B and Leisch F (2007). Applications of finite mixtures of regression models.

APPENDICES

APPENDIX A

Code

A.1 R Function for Fitting a Mixture of Logistic Regressions

```
logistic_regression_EM <- function(
  z,
  m,
  y.known,
  z.known,
  X,
  initial.p,
  initial.beta,
  max.iter = 5000
){
  n.groups <- length(initial.p)
  ### Initial Values
  p <- initial.p
  beta <- initial.beta
  mu <- 1/(1 + exp(-as.vector(X%*%beta)))
  y.unknown <- m - z - apply(y.known, 2, sum)
  z.unknown <- z - apply(z.known, 2, sum)
  i <- 0
  my.loglik <-
    sum(z.unknown*log(apply(p * matrix(mu, nrow = n.groups), 2, sum))) +
    sum((y.known*log(p*(1 - matrix(mu, nrow = n.groups))))[y.known!=0]) +
    lfactorial(sum(m)) -
    sum(lfactorial(sum(z.unknown)), lfactorial(apply(y.known, 1, sum)))
  EM.converged <- F
  starttime <- Sys.time()
  while(!EM.converged){
    i <- i + 1
    cat(i, sep = "\n")

    ### Expectation
    pgivenz <- as.vector(p*matrix(mu, nrow = n.groups))/
      rep(apply(matrix(p*mu, nrow = n.groups), 2, sum), each = n.groups)
    pgiveny <- as.vector(p*(1 - matrix(mu, nrow = n.groups)))/
      rep(
        apply(matrix(p*(1 - mu), nrow = n.groups), 2, sum),
```

```
      each = n.groups
    )
  y.long <- as.vector(y.known) + rep(y.unknown, each = n.groups)*pgiveny
  z.long <- as.vector(z.known) + rep(z.unknown, each = n.groups)*pgivenz
  m.long <- z.long + y.long

  ### Maximization
  fit <- try(
    glm(cbind(z.long, y.long) ~ -1 + X, family = "binomial"),
    silent = TRUE
  )
  if("try-error" %in% class(fit)) {break}
  # Directly for p
  p_prime <- apply(matrix(z.long + y.long, nrow = n.groups), 1, sum) /
  sum(z.long + y.long)
  beta_prime <- coef(fit)
  ll_new <-
    sum(lgamma(m.long + 1) - lgamma(z.long + 1) - lgamma(y.long + 1)) +
    (z.long %*% -log(1 + exp(as.vector(-X%*%beta_prime)))) +
    ((y.long) %*% -log(1 + exp(as.vector(X%*%beta_prime))))

  ### Convergence of EM
  EM.converged <- all(abs(c(p_prime - p, beta_prime - beta)) < 1e-10) |
    i == max.iter

  p <- p_prime
  beta <- beta_prime
  mu <- 1/(1 + exp(-as.vector(X%*%beta)))
  my.ll <- sum(
    z.unknown*log(apply(p * matrix(mu, nrow = n.groups), 2, sum))) +
    sum((y.known*log(p*(1 - matrix(mu, nrow = n.groups))))[y.known!=0]) +
    lfactorial(sum(m)) -
    sum(lfactorial(sum(z.unknown)), lfactorial(apply(y.known, 1, sum))
  )
  my.loglik <- c(my.loglik, my.ll)
}
endtime <- Sys.time()
list(
  my.loglik = my.loglik,
  parameters = matrix(
    c(beta, p),
    nrow = n.groups,
    dimnames = list(
      genotype = 1:n.groups,
      parameter = c("b0", "b1", "p")
    )
```

```
    ),
    unpenalized.loglik = my.ll,
    time = difftime(endtime, starttime, units = "secs"),
    converged = i < max.iter
  )
}
```

A.2 R Function for Fitting a Mixture of Penalized Logistic Regressions

```
firth_logistic_regression_EM <- function(z, m, y.known, z.known, X,
  A = NULL, a = NULL, which.group = NULL, initial.p, initial.beta){
  n.groups <- length(initial.p)
  ### Initial Values
  p <- initial.p
  beta <- initial.beta
  mu <- 1/(1 + exp(-as.vector(X%*%beta)))
  y.unknown <- m - z - apply(y.known, 2, sum)
  z.unknown <- z - apply(z.known, 2, sum)

  pgivenz <- as.vector(p*matrix(mu, nrow = n.groups))/
    rep(apply(matrix(p*mu, nrow = n.groups), 2, sum), each = n.groups)
  pgiveny <- as.vector(p*(1 - matrix(mu, nrow = n.groups)))/
    rep(apply(matrix(p*(1 - mu), nrow = n.groups), 2, sum), each = n.groups)
  y.long <- as.vector(y.known) + rep(y.unknown, each = n.groups)*pgiveny
  z.long <- as.vector(z.known) + rep(z.unknown, each = n.groups)*pgivenz
  m.long <- z.long + y.long

  my.loglik <-
    sum(
      y.unknown*log(apply(p * (1 - matrix(mu, nrow = n.groups)), 2, sum))
    ) +
    sum(z.unknown*log(apply(p * matrix(mu, nrow = n.groups), 2, sum))) +
    sum((y.known*log(p * (1 - matrix(mu, nrow = n.groups))))[y.known!=0]) +
    sum((z.known*log(p * matrix(mu, nrow = n.groups)))[z.known!=0]) +
    lfactorial(sum(m)) -
    sum(
      lfactorial(sum(z.unknown + y.unknown)),
      lfactorial(apply(z.known, 1, sum) + apply(y.known, 1, sum))
    ) + 1/2 * log(det(t(X) %*% diag(m.long/(1 + exp(as.vector(-X%*%beta)))/
      (1 + exp(as.vector(X%*%beta)))) %*% X))
  i <- 0
  EM.converged <- F
  starttime <- Sys.time()
  while(!EM.converged){
    i <- i + 1
```

```
    #cat(i, sep = "\n")

    ### Expectation
    pgivenz <- as.vector(p*matrix(mu, nrow = n.groups))/
      rep(apply(matrix(p*mu, nrow = n.groups), 2, sum), each = n.groups)
    pgiveny <- as.vector(p*(1 - matrix(mu, nrow = n.groups)))/
      rep(apply(matrix(p*(1 - mu), nrow = n.groups), 2, sum), each = n.groups)
    y.long <- as.vector(y.known) + rep(y.unknown, each = n.groups)*pgiveny
    z.long <- as.vector(z.known) + rep(z.unknown, each = n.groups)*pgivenz
    m.long <- z.long + y.long

    ### Maximization
    # Directly for p
    p_new <- apply(matrix(z.long + y.long, nrow = n.groups), 1, sum) /
      sum(z.long + y.long)
    # Newton method for beta
    beta_LR <- beta
    ll <-
      sum(lgamma(m.long + 1) - lgamma(z.long + 1) - lgamma(y.long + 1)) +
        (z.long %*% -log(1 + exp(as.vector(-X%*%beta_LR)))) +
        (y.long %*%-log(1 + exp(as.vector(X%*%beta_LR))))
    LR.converged <- FALSE
    #j <- 0
    while(!LR.converged){
      #j <- j + 1
      #cat(j, sep = "\n")
      eta <- as.vector(X%*%beta_LR)
      mu <- 1/(1 + exp(-eta))
      W <- diag(m.long*mu*(1-mu))
      information <- t(X)%*%W%*%X
#       inv.information <- qr.solve(information)
      inv.information <- solve(information)
      hat <- sqrt(W)%*%X%*%inv.information%*%t(X)%*%sqrt(W)
      score <- t(X)%*%(z.long - m.long*mu + diag(hat)*(1/2 - mu))
      change <- as.vector(inv.information%*%score)
      # Step halving
      tuner <- 1#/2
      ll.increased <- FALSE
      while(!ll.increased){
        beta_new <- beta_LR + tuner*change
        if(!is.null(A)){
          inv.information.p <- diag(p_new^2/apply(matrix(m.long, nrow =
            n.groups), 1, sum))
          inv.information.all <- cbind(rbind(inv.information, matrix(0,
            ncol = ncol(inv.information), nrow = nrow(inv.information.p))),
            rbind(matrix(0, nrow = nrow(inv.information), ncol = ncol(
```

```
        inv.information.p)), inv.information.p))
      parameters_new <- c(beta_new, p_new) + inv.information.all%*%
        t(A)%*%solve(A%*%inv.information.all%*%t(A))%*%(a - A%*%c(
        beta_new, p_new))
      beta_blah <- beta_new
      beta_new <- parameters_new[1:length(beta)]
      p_new <- parameters_new[(length(beta)+1):length(parameters_new)]
      mu_new <- 1/(1 + exp(-as.vector(X%*%beta_new)))
    }
    ll_new <-
      sum(lgamma(m.long + 1) - lgamma(z.long + 1) - lgamma(y.long + 1)) +
      (z.long %*% -log(1 + exp(as.vector(-X%*%beta_new)))) +
      ((y.long) %*% -log(1 + exp(as.vector(X%*%beta_new))))
    ll.increased <- TRUE#ll_new >= ll & !is.na(ll_new)
    tuner <- tuner/2
  }
  # Convergence of newton method
  LR.converged <- all(abs(beta_LR - beta_new) < 1e-5)# & ll_new >= ll
  beta_LR <- beta_new
  ll <- ll_new
}
p_prime <- p_new
beta_prime <- beta_LR

### Convergence of EM
EM.converged <- all(abs(c(p_prime - p, beta_prime - beta)) < 1e-5) |
  i > 5000

p <- p_prime
beta <- beta_prime
mu <- 1/(1 + exp(-as.vector(X%*%beta)))
my.ll <-
  sum(
    y.unknown*log(apply(p * (1 - matrix(mu, nrow = n.groups)), 2, sum))
  ) +
  sum(z.unknown*log(apply(p * matrix(mu, nrow = n.groups), 2, sum))) +
  sum(
    (y.known*log(p * (1 - matrix(mu, nrow = n.groups))))[y.known!=0]
  ) +
  sum((z.known*log(p * matrix(mu, nrow = n.groups)))[z.known!=0]) +
  lfactorial(sum(m)) -
  sum(
    lfactorial(sum(z.unknown + y.unknown)),
    lfactorial(apply(z.known, 1, sum) + apply(y.known, 1, sum))
  ) + 1/2 *
  log(det(
```

```
        t(X) %*%
        diag(
          m.long/
          (1 + exp(as.vector(-X%*%beta)))/
          (1 + exp(as.vector(X%*%beta)))
        ) %*%
        X
      ))
    my.loglik <- c(my.loglik, my.ll)
  }
  endtime <- Sys.time()
  list(
    my.loglik = my.loglik,
    z = z,
    m = m,
    y.known = y.known,
    z.known = z.known,
    y.unknown = y.unknown,
    z.unknown = z.unknown,
    X = X,
    parameters = matrix(
      c(beta, p),
      nrow = n.groups,
      dimnames = list(genotype = 1:n.groups, parameter = c("b0", "b1", "p"))
    ),
    penalized.loglik = my.ll,
    time = difftime(endtime, starttime, units = "secs")
  )
}
```

A.3 R Function For Profiling

```
profile <- function(
  fit,
  parameter = c("b0", "b1", "p", "prediction", "inverse"),
  group,
  upper = TRUE,
  pred.pi = NULL,
  step = NA
){
  z <- fit$z
  m <- fit$m
  y.known <- fit$y.known
  z.known <- fit$z.known
  X <- fit$X
```

```
n.groups <- nrow(fit$parameters)

if(parameter == "prediction"){
  mle <- pred.pi
  A <- matrix(0, nrow = 2, ncol = 3*n.groups)
  A[1, group + n.groups*(which(c("b0", "b1", "p") == "b0") - 1)] <- 1
  A[1, group + n.groups*(which(c("b0", "b1", "p") == "b1") - 1)] <-
    (log(pred.pi/(1-pred.pi))-fit$parameters[group, "b0"])/
    fit$parameters[group, "b1"]
  A[2, (n.groups*2+1):(n.groups*3)] <- 1
} else {
  if(parameter == "inverse"){
    mle <- (log(pred.pi/(1-pred.pi)) - fit$parameters[group, "b0"])/
      fit$parameters[group, "b1"]
    a <- c(log(pred.pi/(1-pred.pi)), 1)
    A <- matrix(0, nrow = 2, ncol = 3*n.groups)
    A[1, group + n.groups*(which(c("b0", "b1", "p") == "b0") - 1)] <- 1
    A[2, (n.groups*2+1):(n.groups*3)] <- 1
  } else {
    mle <- fit$parameters[group, parameter]
    A <- matrix(0, nrow = 2, ncol = 3*n.groups)
    A[1,
      group + n.groups*(which(c("b0", "b1", "p") == parameter) - 1)] <- 1
    A[2, (n.groups*2+1):(n.groups*3)] <- 1
  }
}

if(parameter %in% c("p", "prediction")){
  old.test <- round(log(mle/(1-mle)), 1)
} else {
  old.test <- mle
}

old.fit <- fit

if(parameter %in% c("p", "prediction", "inverse")){
  old.test <- ifelse(
    upper,
    floor(old.test*10)/10,
    ceiling(old.test*10)/10
  )
}

profile.loglik <- NULL
l <- 0
```

```
  while(old.fit$penalized.loglik > (fit$penalized.loglik - 4) &
    !(parameter == "prediction" & abs(old.test) > 6)){
    l <- l + 1
    cat(l, sep = "\n")
    #     new.test <- old.test + step
    new.test <- old.test + step*(2*upper-1)
    if(parameter == "inverse"){
      A[1, group + n.groups*(which(c("b0", "b1", "p") == "b1") - 1)] <-
        new.test
    } else {
      if(parameter == "p"){
        a <- c(1/(1+exp(-new.test)), 1)
      } else {
        a <- c(new.test, 1)
      }
    }
    parameters.start <- old.fit$parameters
    new.fit <- firth_logistic_regression_EM(
      z = z,
      m = m,
      y.known = y.known,
      z.known = z.known,
      X = X,
      A = A,
      a = a,
      initial.p = as.vector(parameters.start[, "p"]),
      initial.beta = as.vector(parameters.start[, c("b0", "b1")])
    )
    profile.loglik <- rbind(
      profile.loglik,
      c(new.test, new.fit$penalized.loglik, as.vector(new.fit$parameters))
    )
    old.fit <- new.fit
    old.test <- new.test
  }
  profile.loglik
}
```

A.4 R Function for Finding Profile Likelihood Interval Endpoints

```
locate_endpoint <- function(
  fit,
  parameter = c("b0", "b1", "p", "prediction", "inverse"),
  group,
```

```
  upper = TRUE,
  pred.pi = NULL,
  step = NA,
  alpha = 0.05
){
  z <- fit$z
  m <- fit$m
  y.known <- fit$y.known
  z.known <- fit$z.known
  X <- fit$X

  n.groups <- nrow(fit$parameters)

  if(parameter == "prediction"){
    mle <- pred.pi
    A <- matrix(0, nrow = 2, ncol = 3*n.groups)
    A[1, group + n.groups*(which(c("b0", "b1", "p") == "b0") - 1)] <- 1
    A[1, group + n.groups*(which(c("b0", "b1", "p") == "b1") - 1)] <-
      (log(pred.pi/(1-pred.pi))-fit$parameters[group, "b0"])/
      fit$parameters[group, "b1"]
    A[2, (n.groups*2+1):(n.groups*3)] <- 1
  } else {
    if(parameter == "inverse"){
      mle <- (log(pred.pi/(1-pred.pi)) - fit$parameters[group, "b0"])/
        fit$parameters[group, "b1"]
      a <- c(log(pred.pi/(1-pred.pi)), 1)
      A <- matrix(0, nrow = 2, ncol = 3*n.groups)
      A[1, group + n.groups*(which(c("b0", "b1", "p") == "b0") - 1)] <- 1
      A[2, (n.groups*2+1):(n.groups*3)] <- 1
    } else {
      mle <- fit$parameters[group, parameter]
      A <- matrix(0, nrow = 2, ncol = 3*n.groups)
      A[1,
        group + n.groups*(which(c("b0", "b1", "p") == parameter) - 1)] <- 1
      A[2, (n.groups*2+1):(n.groups*3)] <- 1
    }
  }

  new.fit <- fit

  if(parameter %in% c("p", "prediction")){
    new.test <- log(mle/(1-mle))
  } else {
    new.test <- mle
  }
```

```
  profile.loglik <- NULL
  converged <- FALSE
  beyond.threshold <- FALSE
  i <- 0
  while(!converged){
    i <- i + 1
#     cat(i, sep = "\n")
    if(!beyond.threshold){
      high.fit <- new.fit
      high.test <- new.test
      new.test <- high.test + step*(2*upper-1)
      parameters.start <- new.fit$parameters
    } else {
      if(new.fit$penalized.loglik >
        (fit$penalized.loglik - qchisq(1-alpha,1)/2)){
        high.fit <- new.fit
        high.test <- new.test
      } else {
        low.fit <- new.fit
        low.test <- new.test
      }
      new.test <- mean(c(low.test, high.test))
      parameters.start <- high.fit$parameters
    }
    if(parameter == "inverse"){
      A[1, group + n.groups*(which(c("b0", "b1", "p") == "b1") - 1)] <-
        new.test
    } else {
      if(parameter == "p"){
        a <- c(1/(1+exp(-new.test)), 1)
      } else {
        a <- c(new.test, 1)
      }
    }
    new.fit <- firth_logistic_regression_EM(
      z = z,
      m = m,
      y.known = y.known,
      z.known = z.known,
      X = X,
      A = A,
      a = a,
      initial.p = as.vector(parameters.start[, "p"]),
      initial.beta = as.vector(parameters.start[, c("b0", "b1")])
    )
    if(!beyond.threshold){
```

```
      beyond.threshold <- new.fit$penalized.loglik <
        (fit$penalized.loglik - qchisq(0.95, 1)/2)
    } else {
      converged <- beyond.threshold & abs(low.test - high.test) < 1e-5
    }
  }
  new.test
}
```

A.5 Example Application of the Previous Functions to the *Rhyzopertha dominica* Data

```
###
### Schlipalius et al 2002
###
m <- c(98, 100, 100, 100, 100, 300, 400, 750, 500, 500, 7850)

dose <- c(0, 0.003, 0.004, 0.005, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 1)
logdose <- log(0.004 + dose)

logdose.long <- rep(logdose, each = 6)
genotype <- factor(
  rep(
    paste(
      rep(c("-", "+"), each=3),
      rep(c("B", "H", "A"), 2), sep="/"
    ),
    times = 11
  ),
  levels = paste(
    rep(c("-", "+"), each=3),
    rep(c("B", "H", "A"), 2),
    sep="/"
  )
)

X <- data.matrix(model.matrix(~-1 + genotype + genotype:logdose.long))
genotype.labels <- paste(
  rep(c("-", "+"), each=3),
  rep(c("B", "H", "A"), 2),
  sep="/"
)
colnames(X) <- paste(
  rep(genotype.labels, 2),
  rep(c("b0", "b1"), each = 6),
  sep = "_"
```

```
)

genotype.key <- data.frame(
  group = 1:6,
  genotype = genotype.labels,
  rp6.79 = rep(c("$-$", "$+$"), each = 3),
  rp5.11 = rep(c("B", "H", "A"), times = 2),
  stringsAsFactors = FALSE
)

y <- matrix(nrow = 6, ncol = 11)
y[1,] <- c(31, 18, 10, 1, 0, 0, 0, 0, 0, 0, 0)
y[2,] <- c(27, 26, 4, 4, 1, 0, 0, 0, 0, 0, 0)
y[3,] <- c(10, 10, 3, 7, 9, 0, 0, 0, 0, 0, 0)
y[4,] <- c(6, 6, 4, 2, 8, 5, 0, 0, 0, 0, 0)
y[5,] <- c(20, 20, 7, 6, 5, 20, 10, 0, 0, 0, 0)
y[6,] <- c(4, 4, 4, 2, 0, 5, 7, 10, 10, 8, 44)

### Run EM
fit <- firth_logistic_regression_EM(
  z = m - apply(y, 2, sum),
  m = m,
  y.known = y,
  z.known = matrix(0, nrow = 6, ncol = 11),
  X = X,
  initial.p = rep(1/6, 6),
  initial.beta = rep(0, 12)
)

### Profile $\beta_0$ above its point estimate
profile(fit,
  parameter = "b0",
  group = 1,
  upper = TRUE,
  step = 2
)

### Extract lower confidence limit for $LD_{99}$
locate_endpoint(fit,
  parameter = "inverse",
  group = i,
  upper = FALSE,
  pred.pi = 0.99,
  alpha = 0.05,
  step = 0.1
)
```

A.6 Function for Simulating from Mixture of Dose-Response Curves

```
simulate_mixture <- function(parameters, logdose, samplesize) {
  d <- length(logdose)
  # m <- rep(samplesize, d)
  m <- samplesize

  a.b0 <- parameters[1, 1]
  a.b1 <- parameters[1, 2]
  b.b0 <- parameters[2, 1]
  b.b1 <- parameters[2, 2]

  p <- parameters[1, 3]

  genotype.a <- rbinom(d, m, p)
  genotype.b <- m - genotype.a

  z.a <- rbinom(d, genotype.a, 1/(1 + exp(-(a.b0 + logdose * a.b1))))
  z.b <- rbinom(d, genotype.b, 1/(1 + exp(-(b.b0 + logdose * b.b1))))

  logdose.long <- rep(logdose, each = 2)
  genotype.labels <- c("a", "b")

  genotype <- factor(rep(genotype.labels, times = d), levels =
    genotype.labels)

  X <- data.matrix(model.matrix(~-1 + genotype + genotype:logdose.long))
  colnames(X) <- paste(rep(genotype.labels, 2), rep(c("b0", "b1"), each =
    2), sep = "_")

  list(
    parameters = parameters,
    logdose = logdose,
    m = m,
    genotype = rbind(genotype.a, genotype.b),
    z = rbind(z.a, z.b),
    X = X
  )
}
```

A.7 Code for Labeling and Fitting Models to Simulated Data

```
# Simulate Data
x <- simulate_mixture(
  parameters = cbind(fit$parameters[unlist(setup[i, c("low", "high")]),
```

```
    1:2],
  p = c(0.5, 0.5)),
  logdose = log(1/2^(9:0)),
  samplesize = rep(100, times = 10)
)

# Label all live beetles and fit unpenalized model
logistic_regression_EM(
  z = apply(x$z, 2, sum),
  m = x$m,
  y.known = x$genotype - x$z,
  z.known = matrix(0, nrow = 2, ncol = length(x$m)),
  X = x$X,
  initial.p = x$parameters[, "p"],
  initial.beta = as.vector(x$parameters[, c("b0", "b1")]),
  max.iter = 1000
)

# Label all live beetles and fit penalized model
firth_logistic_regression_EM(
  z = apply(x$z, 2, sum),
  m = x$m,
  y.known = x$genotype - x$z,
  z.known = matrix(0, nrow = 2, ncol = length(x$m)),
  X = x$X,
  initial.p = x$parameters[, "p"],
  initial.beta = c(0, 0, 0, 0)
)

# Label random dead and live beetles
y <- x$m - apply(x$z, 2, sum)
all <- rbind(x$z, x$genotype - x$z)
known <- NULL
for(i in 1:length(y)){
  known <- cbind(
    known,
    table(sample(rep(factor(1:4), times = all[, i]), y[i]))
  )
}
z.known <- known[1:2, ]
y.known <- known[3:4, ]

#Fit unpenalized model
logistic_regression_EM(
  z = apply(x$z, 2, sum),
  m = x$m,
```

```
  y.known = y.known,
  z.known = z.known,
  X = x$X,
  initial.p = x$parameters[, "p"],
  initial.beta = as.vector(x$parameters[, c("b0", "b1")]),
  max.iter = 1000
)

#Fit penalized model
firth_logistic_regression_EM(
  z = apply(x$z, 2, sum),
  m = x$m,
  y.known = y.known,
  z.known = z.known,
  X = x$X,
  initial.p = x$parameters[, "p"],
  initial.beta = c(0, 0, 0, 0)
)
```

CURRICULUM VITAE

Darl D. Flake II

PhD in Mathematical Sciences, Utah State University, 2016.

MS in Statistics, Utah State University, 2008.

BS in Interdisciplinary Studies, Utah State University, 2006.

Research interests include applied biostatistics.